PROGRAMME AND ABSTRACTS

13th International Conference on

Computational and Financial Econometrics (CFE 2019)

http://www.cfenetwork.org/CFE2019

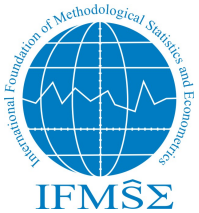
and

12th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2019)

http://www.cmstatistics.org/CMStatistics2019

Senate House & Birkbeck University of London, UK
14 – 16 December 2019



ISBN 978-9963-2227-8-0

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes and Manfred Deistler.

CFE 2019 Co-chairs:
Josu Arteche, Walter Distaso, Roxana Halbleib and Mathieu Rosenbaum.

CFE 2019 Programme Committee:

Alessandra Amendola, Jonas Andersson, Jozef Barunik, Andrea Berardi, Scott Brave, Wojciech Charemza, Luca De Angelis, Tomas Del Barrio, Peter Exterkate, Marzia Freo, Luis Alberiko Gil Alana, Benjamin Holcblat, Yuan Ke, Robert Kohn, Robinson Kruse-Becher, Degui Li, Loriano Mancini, Federico Martellosio, Simone Maxand, Nour Meddahi, Antonio Montanes, Gabriel Montes-Rojas, Jose Olmo, Michael Owyang, Laurent Pauwels, Peter Pedroni, Eric Renault, Marialuisa Restaino, Paulo M.M. Rodrigues, Leopold Soegner, Marco Maria Sorge, James Taylor, Nora Traum, Bezirgen Veliyev, Helga Wagner, Claudia Wellenreuther, Yohei Yamamoto and Wenying Yao.

CMStatistics 2019 Co-chairs:
Jochen Einbeck, Michael Daniels, Chenlei Leng, Peter Rousseeuw and Grace Yi.

CMStatistics 2019 Programme Committee:

Mihye Ahn, Jose Ameijeiras-Alonso, Andreas Anastasiou, Yuko Araki, Sergio Bacallado, Eric Beutner, Paula Bouzas, Babette Brumback, Juan Juan Cai, Eva Cantoni, Yuguo Chen, Sy Han Chiou, Bertrand Clarke, Radu Craiu, Federico Crudu, Hongsheng Dai, Matthew Davison, Miguel De Carvalho, F Marta L Di Lascio, Takeshi Emura, Davide Ferrari, M. Brigida Ferraro, Marco Ferreira, Yulia Gel, Subir Ghosh, Jeff Goldsmith, Gil Gonzalez-Rodriguez, Andreas Groll, Alessandra Guglielmi, Sebastien Haneuse, Wenqing He, Jennifer Hill, Raphael Huser, Marie Huskova, Piotr Jaworski, Binyan Jiang, Timothy Johnson, Linglong Kong, Efoevi Angelo Koudou, Johannes Lederer, Xiaodong Li, Nicola Loperfido, Sara Lopez Pintado, Monia Lupparelli, Shujie Ma, Taps Maiti, Hiroki Masuda, Andreas Mayr, Hans Mueller, Bernardo Nipoti, Klaus Nordhausen, Marco Oesting, Guangming Pan, Wei Pan, Juhyun Park, Tetyana Pavlenko, Mario Peruggia, Zuzana Praskova, Anne Ruiz-Gazen, Sylvain Sardy, Yuya Sasaki, Adrien Saumard, Michael Schweinberger, Russell Shinohara, Rosaria Simone, Ekaterina Smirnova, Xinyuan Song, Emmanuele Sordini, Natalia A Stepanova, Cheng Yong Tang, Sara Taskinen, Wolfgang Trutschnig, Germain Van Bever, Cristiano Varin, Tim Verdonck, Anneleen Verhasselt, Mattias Villani, HaiYing Wang, Ines Wilms, Changbao Wu, Lang Wu, Min-ge Xie, Xinyi Xu, Nakahiro Yoshida, Yi Yu, Xin Zhang, Zhigen Zhao, Jin Zhou, Ji Zhu and Yunzhang Zhu.

Local Organizer:
Birkbeck University of London.
CFEnetwork and CMStatistics.

Dear Friends and Colleagues,

We welcome you warmly to London, for the 13th International Conference on *Computational and Financial Econometrics* (CFE 2019) and the 12th International Conference of the ERCIM Working Group on *Computational and Methodological Statistics* (CMStatistics 2019). As many of you know, this annual conference has become a leading joint international meeting at the interface of statistics, econometrics, empirical finance and computing.

The conference aims at bringing together researchers and practitioners to discuss recent developments in computational methods for economics, finance, and statistics. The CFE-CMStatistics 2019 programme consists of 430 sessions, 5 plenary talks and over 1720 presentations. There are about 1880 participants. Once more, this is the biggest meeting of the conference series in terms of number of participants and presentations. The growth of the conference in terms of size and quality makes it undoubtedly one of the most important international scientific events in the field.

The co-chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. The international organizing committee hopes that the conference venue will provide the appropriate environment to enhance your contacts and to establish new ones. The conference is a collective effort by many individuals and organizations. The Scientific Programme Committee, the Session Organizers, the local hosting universities and many volunteers have contributed substantially to the organization of the conference. We acknowledge their work and the support of our hosts and sponsors, and particularly, Birkbeck University of London, UK.

The Elsevier journal, Econometrics and Statistics (EcoSta), has being inaugurated in 2017. The EcoSta is the official journal of the networks of Computational and Financial Econometrics (CFEnetwork) and of Computational and Methodological Statistics (CMStatistics). It publishes research papers in all aspects of econometrics and statistics and it comprises two sections, namely, Part A: Econometrics and Part B: Statistics. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta and its supplement Annals of Computational and Financial Econometrics.

The CMStatistics will commence The Annals of Statistical Data Science (SDS) which will be published as a supplement of the Elsevier journal Computational Statistics & Data Analysis (CSDA). The CSDA is also the official journal of CMStatistics. You are encouraged to submit your papers to the Annals of Statistical Data Science or regular peer-reviewed issues of CSDA.

Looking forward, the CFE-CMStatistics 2020 will be held at King's College London, from Saturday the 19th of December 2020 to Monday the 21st of December 2020. Tutorials will take place on Thursday the 18th of December 2020. You are invited and encouraged to actively participate in these events.

We wish you a productive, stimulating conference and a memorable stay in London.

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler
Coordinators of CMStatistics & CFEnetwork and EcoSta.

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

The working group (WG) CMStatistics comprises a number of specialized teams in various research areas of computational and methodological statistics. The teams act autonomously within the framework of the WG in order to promote their own research agenda. Their activities are endorsed by the WG. They submit research proposals, organize sessions, tracks and tutorials during the annual WG meetings and edit journal special issues. The Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are the official journals of the CMStatistics.

Specialized teams

Currently the ERCIM WG has over 1750 members and the following specialized teams

BM: Bayesian Methodology
CODA: Complex data structures and Object Data Analysis
CPEP: Component-based methods for Predictive and Exploratory Path modeling
DMC: Dependence Models and Copulas
DOE: Design Of Experiments
EF: Econometrics and Finance
GCS: General Computational Statistics WG CMStatistics
GMS: General Methodological Statistics WG CMStatistics
GOF: Goodness-of-Fit and Change-Point Problems
HDS: High-Dimensional Statistics
ISDA: Imprecision in Statistical Data Analysis
LVSEM: Latent Variable and Structural Equation Models
MCS: Matrix Computations and Statistics
MM: Mixture Models
MSW: Multi-Set and multi-Way models
NPS: Non-Parametric Statistics
OHEM: Optimization Heuristics in Estimation and Modelling
RACDS: Robust Analysis of Complex Data Sets
SAE: Small Area Estimation
SAET: Statistical Analysis of Event Times
SAS: Statistical Algorithms and Software
SEA: Statistics of Extremes and Applications
SFD: Statistics for Functional Data
SL: Statistical Learning
SSEF: Statistical Signal Extraction and Filtering
TSMC: Times Series Modelling and Computation

You are encouraged to become a member of the WG. For further information please contact the Chairs of the specialized groups (see the WG's website), or by email at info@cmstatistics.org.

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

The Computational and Financial Econometrics (CFEnetwork) comprises a number of specialized teams in various research areas of theoretical and applied econometrics, financial econometrics and computation, and empirical finance. The teams contribute to the activities of the network by organizing sessions, tracks and tutorials during the annual CFEnetwork meetings, and by submitting research proposals. Furthermore the teams edit special issues currently published under the Annals of CFE. The Econometrics and Statistics (EcoSta) is the official journal of the CFEnetwork.

Specialized teams

Currently the CFEnetwork has over 1000 members and the following specialized teams

AE: Applied Econometrics
BE: Bayesian Econometrics
BM: Bootstrap Methods
CE: Computational Econometrics
ET: Econometric Theory
FA: Financial Applications
FE: Financial Econometrics
TSE: Time Series Econometrics

You are encouraged to become a member of the CFEnetwork. For further information please see the website or contact by email at info@cfenetwork.org.

Schedule

2019-12-14

Opening, 08:25

A - Keynote
CMStatistics
08:40 - 09:30

B
CMStatistics
09:40 - 10:55

Opening, 09:50

C - Keynote
CFE
10:05 - 10:55

Coffee Break
10:55 - 11:25

D
CFE - CMStatistics
11:25 - 13:05

Lunch Break
13:05 - 14:35

E
CFE - CMStatistics
14:35 - 16:15

Coffee Break
16:15 - 16:45

F
CFE - CMStatistics
16:45 - 18:50

Welcome Reception
19:00 - 20:30

2019-12-15

G
CFE - CMStatistics
08:40 - 10:20

Coffee Break
10:20 - 10:50

H
CFE - CMStatistics
10:50 - 12:55

Lunch Break
12:55 - 14:25

I
CFE - CMStatistics
14:25 - 16:05

Coffee Break
16:05 - 16:35

J
CFE - CMStatistics
16:35 - 18:15

K - Keynote
CMStatistics
18:25 - 19:15

L - Keynote
CFE
18:25 - 19:15

Christmas Conference Dinner
20:30 - 23:30

2019-12-16

M
CFE - CMStatistics
08:40 - 10:20

Coffee Break
10:20 - 10:50

N
CFE - CMStatistics
10:50 - 12:55

Lunch Break
12:55 - 14:25

O
CFE - CMStatistics
14:25 - 16:05

Coffee Break
16:05 - 16:35

P
CFE - CMStatistics
16:35 - 17:50

Q - Keynote
CFE - CMStatistics
18:05 - 18:55

Closing, 19:00 - 19:10

TUTORIALS, MEETINGS AND SOCIAL EVENTS

TUTORIALS

Tutorials on "R Programming and mixture models, with application to image analysis" will take place on Friday the 13th of December 2019 at the Senate Room of the Senate House (see maps on pages VIII to X). The tutorials will be delivered by Prof. Jochen Einbeck, Durham University, UK. The first tutorial ("Introduction to R Programming and EM algorithm for Gaussian mixtures") will be 9:00-13:30 and the second one ("Clustering methods for Image Analysis") will be 15:00 to 19:30.

SPECIAL MEETINGS by invitation to group members

- The *CSDA and Annals of Statistical Data Science Editorial Board* and the *Econometrics and Statistics (EcoSta) Editorial Board* meetings will take place on Friday 13th of December 2019, 19:45-20:05 at the Senate room of the Senate House (see maps on pages VIII to X). Before the meetings, from 19:00 to 19:45, there will be a welcome drink at the top of the ceremonial stairs. The CSDA and Annals of Statistical Data Science, and the Ecosta dinner will take place on Friday the 13th of December 2019 at 20:30.
- The *CFE-CMStatistics session organizers* get-together will take place on Friday 13th of December 2019 from 19:00 to 20:00 at the top of the ceremonial stairs of the Senate House.

REGISTRATION AND SOCIAL EVENTS

- The registration will be open on Friday 13th December 2019, 08:30-20:00, at the top of ceremonial stairs of the Senate House. From Saturday 14th of December to Monday 16th of December the registration will be at the McMillan Hall of the Senate House. On Saturday, it will be open from 07:45 to 18:00, on Sunday, from 08:00 to 18:30, and on Monday, from 08:00 to 18:00.
- *The coffee breaks* will take place at the Crush Hall and MacMillan Hall of the Senate House, and at the Foyer of Clore (see maps on pages VIII and IX). You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Saturday 14th of December 2019, from 19:00-20:30.* The Welcome Reception is open to all registrants and accompanying persons who have purchased a reception ticket. It will take place at the Senate House Crush Hall and the MacMillan Hall (see maps on pages VIII and IX). Conference registrants must bring their conference badge in order to attend the reception. Information about the welcome reception booking is embedded in the QR code on the conference badge. Preregistration is required due to health and safety reasons, and the limited capacity of the venue. Entrance to the reception venue will be strictly allowed only to those who have prebooked.
- *Christmas conference Dinner, Sunday 15th of December 2019, from 20:30 to 23:30.* The Christmas Conference Dinner will take place at the Ambassador Bloomsbury Hotel (12 Upper Woburn Pl, Bloomsbury, London WC1H 0HX, see map on page VIII). The conference dinner is optional and registration is required. Participants must bring their conference badge in order to to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.

GENERAL INFORMATION

Addresses of venues

- Birkbeck and Clore Management Centre, University of London, Malet Street, London WC1E 7HX.
- University of London, Senate House, Malet Street, London WC1E 7HU.

Lecture rooms (see maps on pages VIII to XVI)

The paper presentations will take place at Birkbeck, Clore and Senate House. Floor maps of the Senate House and Birkbeck are available on pages IX to XVI. All the rooms in the upper levels of Birkbeck are located in the extension, and they are accesible only from the Lifts B (see map on page XVI). There are no floor maps of Clore, which is a small building and the rooms are easy to find. Due to health and safety regulations the maximum capacity of the rooms should be respected (see the interactive programme for the details). There will be no signs indicating the location of the lecture rooms, and therefore we advise you to visit the venue in advance.

Presentation instructions

The lecture rooms will be equipped with a PC and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide the session chair with the files for the presentation in PDF (Acrobat) on a USB memory stick. This must be done at least ten minutes before each session. Chairs are requested to keep the sessions on schedule. Papers should be presented in the order they are listed in the programme for the convenience of attendees who may wish to go to other rooms mid-session to hear particular papers. In the case of a presenter not attending, please use the extra time for a break or a discussion so that the remaining papers stay on schedule. The PC in the lecture rooms should be used for presentations. An IT technician will be available during the conference and should be contacted in case of problems.

Posters

The poster sessions will take place at the McMillan Hall, Senate House. The posters should be displayed only during their assigned session. The authors are responsible for placing the posters on the panels and removing them after the session. The maximum size of the poster is A0 portrait.

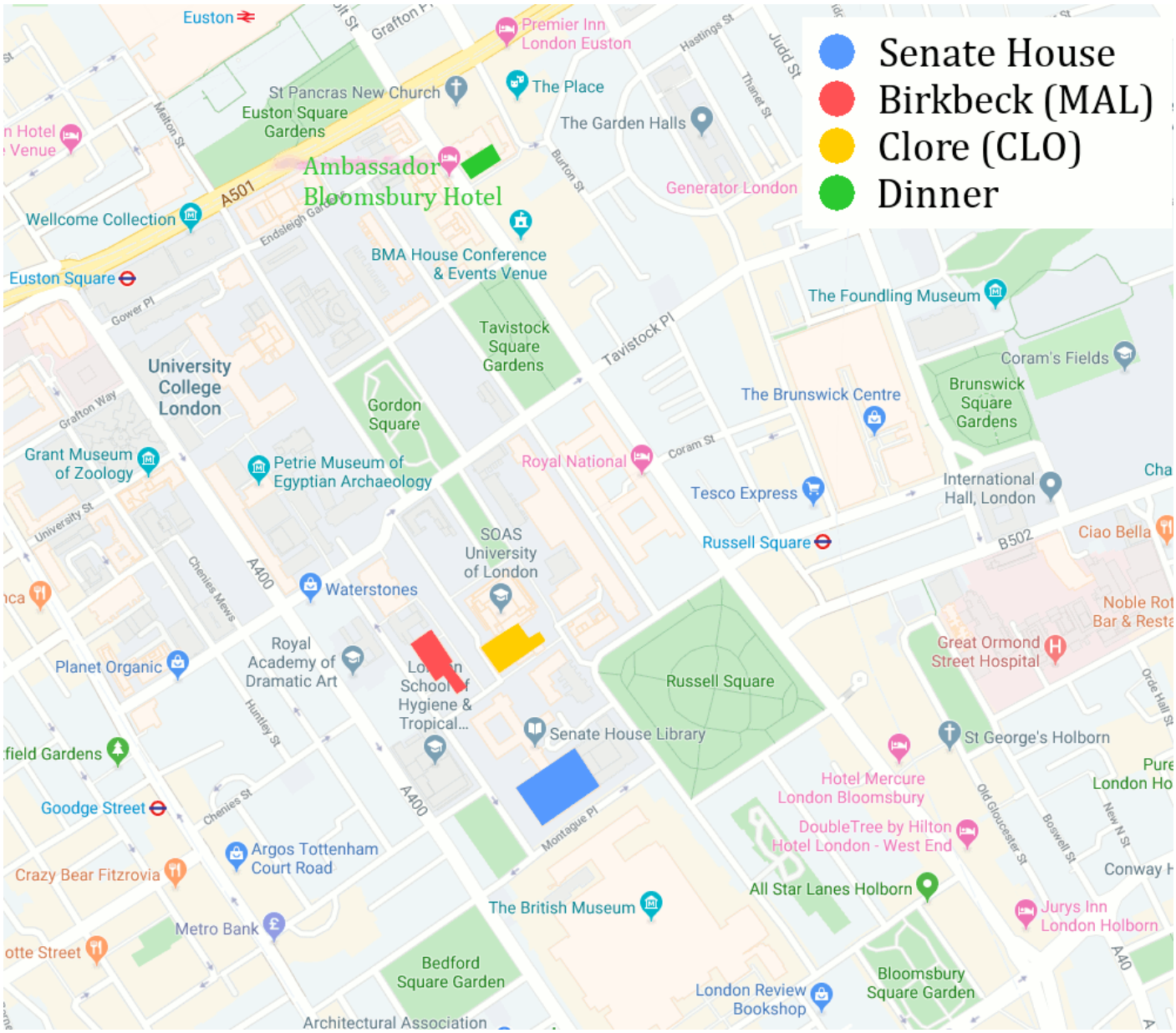
Internet Connection

Participants from any eduroam-enabled institution should use the Eduroam service in order to obtain access to Internet. For participants without Eduroam access, there will be wireless Internet connection at the Macmillan Hall. You will need to have your own laptop in order to connect to the Internet. The daily login and password will be displayed on the announcement board by the registration desk.

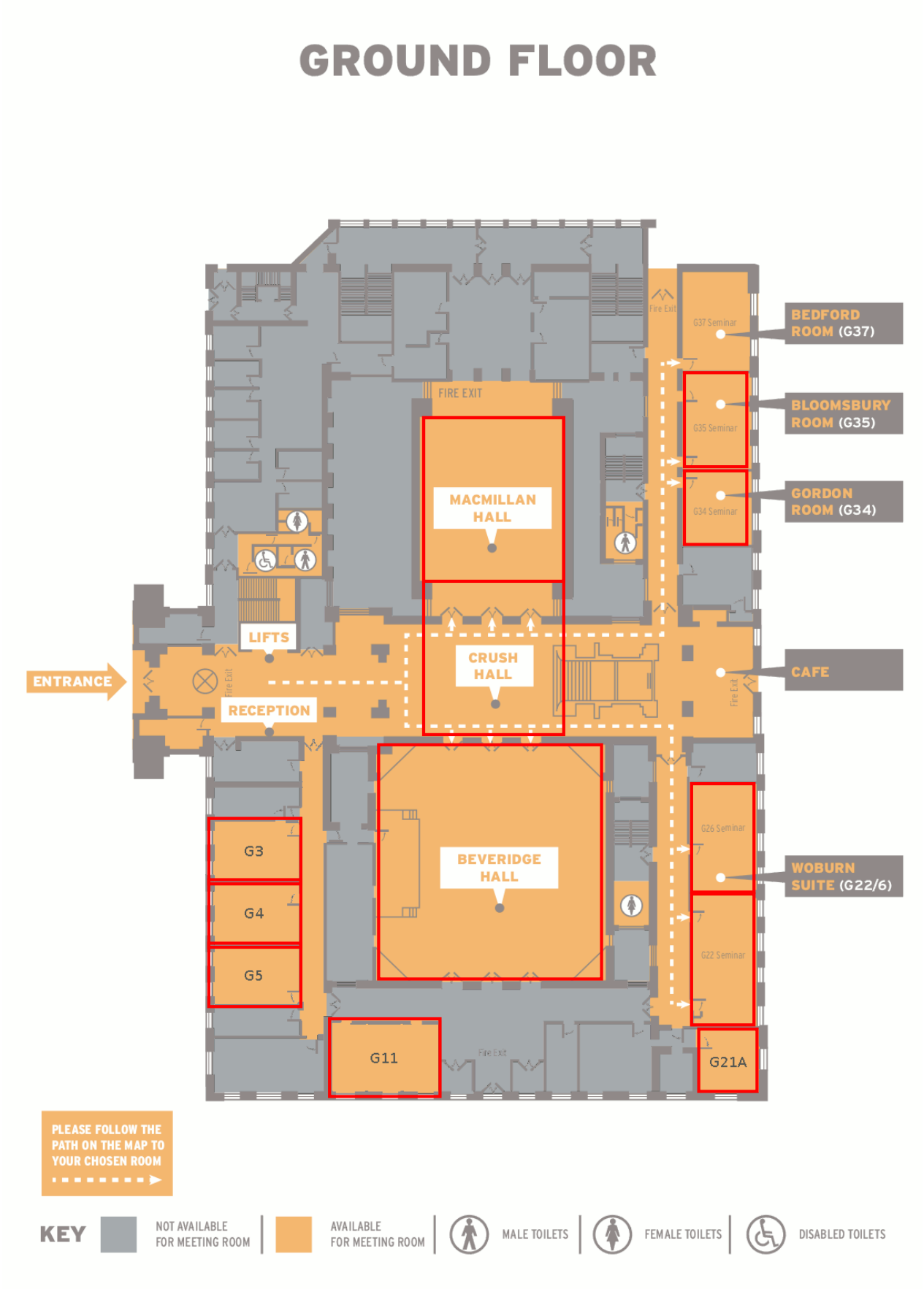
Exhibitors

Elsevier and Springer.

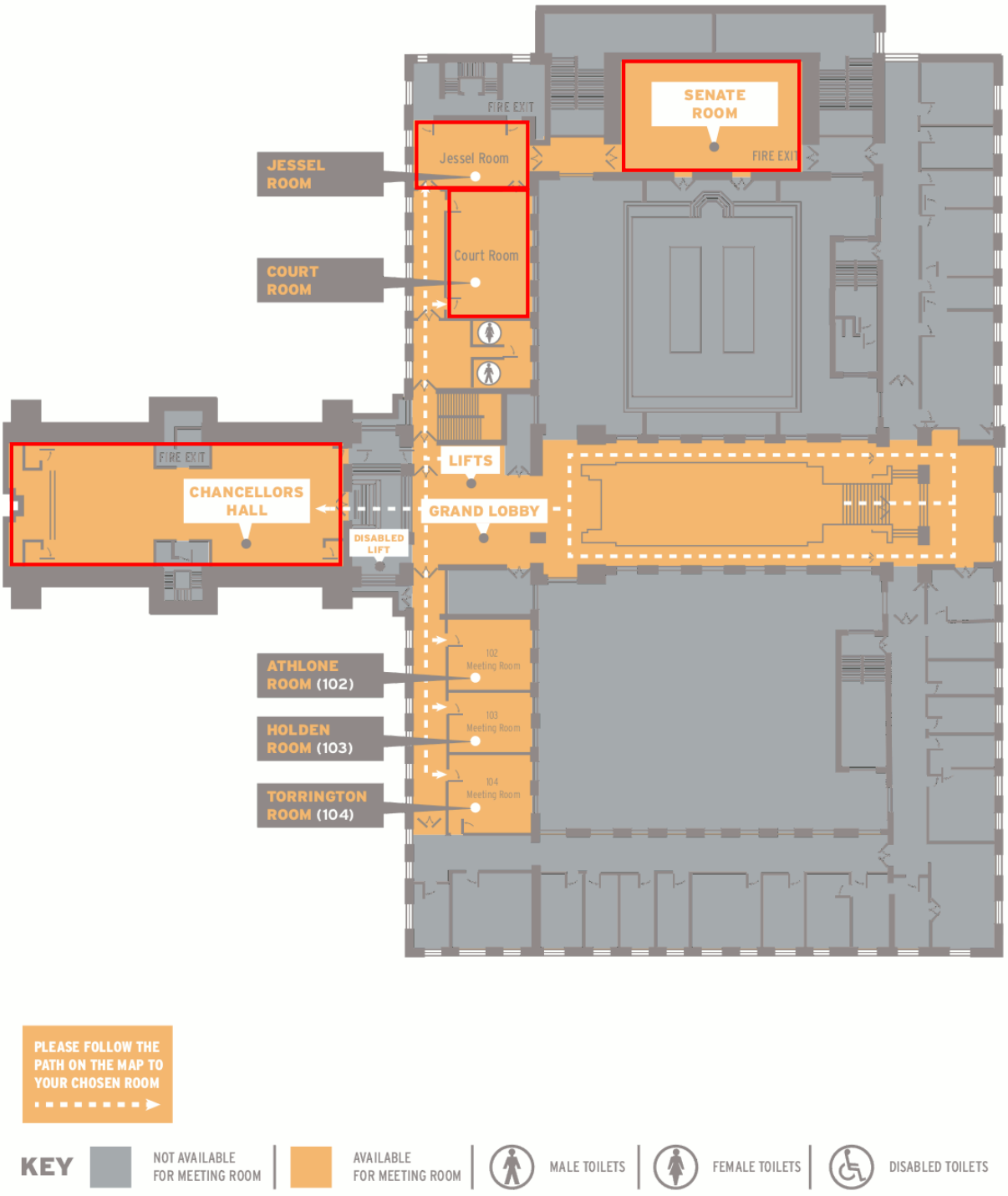
Map of the venue and nearby area

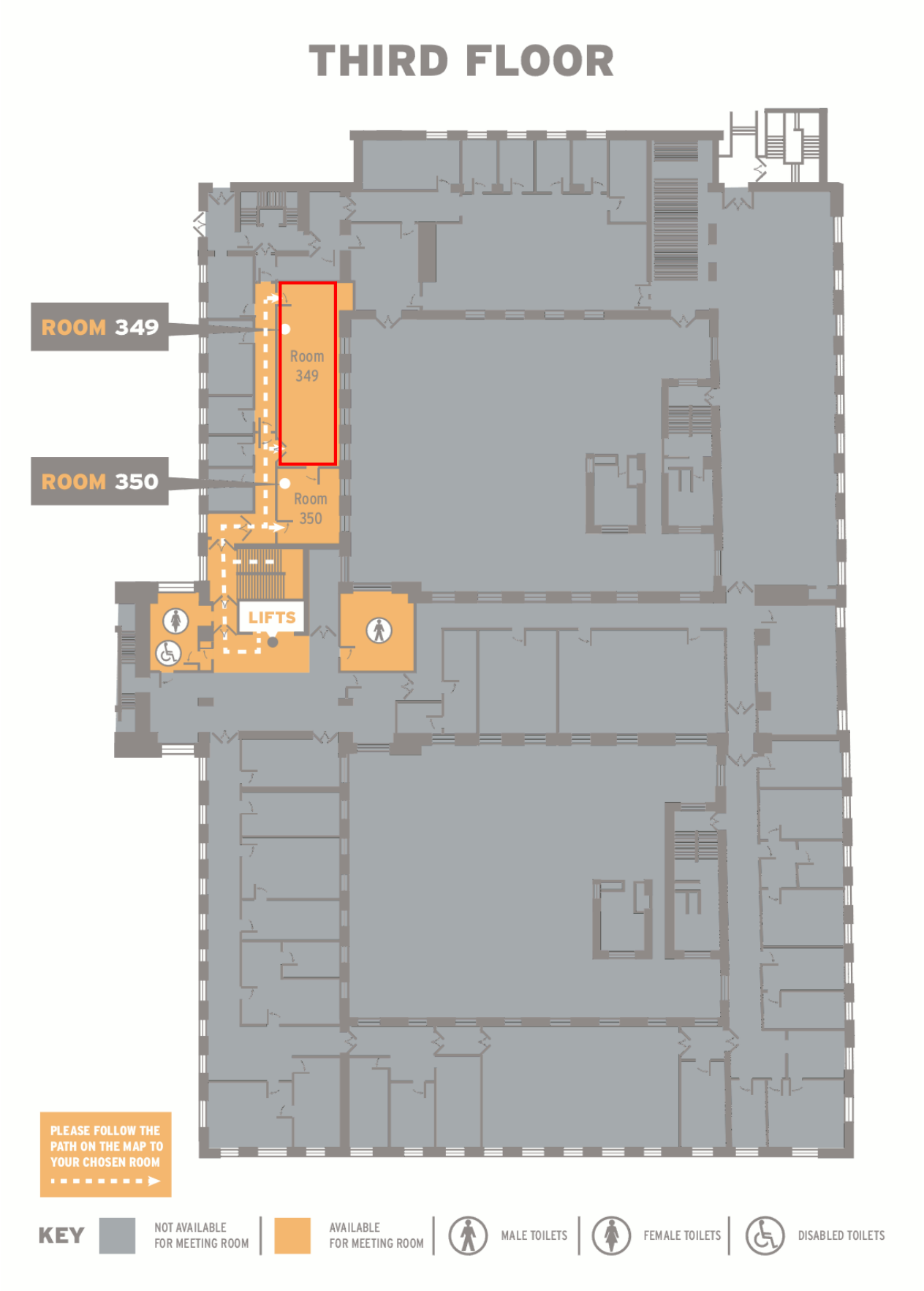


Floor maps – Senate House



FIRST FLOOR





Floor maps – Birkbeck

Key:

New Room Numbers

Old Room Numbers

Revieion A

12 August 2003

Extension Building
Second Floor
(Same level as Mezzanine)
New Room Numbering

ebdi

Access to the rooms in upper levels in Birkbeck

PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics. It publishes research papers in all aspects of econometrics and statistics and comprises two sections:
Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.
Part B: Statistics. Papers providing important original contributions to methodological statistics inspired in applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.
The journal consists, preponderantly, of original research. Occasionally, review and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Papers presented at the conference and containing novel components in econometrics or statistics are encouraged to be submitted for publication in special peer-reviewed or regular issues of the Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. Papers should be submitted using the EM Submission tool. In the EM please select as type of article the CFE conference, CMStatistics Conference or Annals of Computational and Financial Econometrics. Any questions may be directed via email to editor@econometricsandstatistics.org

Call For Papers CSDA Annals of Statistical Data Science (SDS)

http://www.elsevier.com/locate/csda

We are inviting submissions for the 1st issue of the CSDA Annals of Statistical Data Science. The Annals of Statistical Data Science is published as a supplement to the journal of Computational Statistics & Data Analysis. It will serve as an outlet for distinguished research papers using advanced computational and/or statistical methods for tackling challenging data analytic problems. The Annals will become a valuable resource for well-founded theretical and applied data-driven research. Authors submitting a paper to CSDA may request that it be considered for inclusion in the Annals. Each issue will be assigned to several Guest Associate Editors who will be responsible, together with the CSDA Co-Editors, for the selection of papers.

Submissions for the Annals should contain a significant computational or statistical methodological component for data analytics. In particular, the Annals welcomes contributions at the interface of computing, statistics addressing problems involving large and/or complex data. Emphasis will be given to comprehensive and reproducible research, including data-driven methodology, algorithms and software. There is no deadline for submissions. Papers can be submitted at any time. When they have been received, they will enter the editorial system immediately. All submissions must contain original unpublished work not being considered for publication elsewhere. Please submit your paper electronically using the Elsevier Editorial System: http://ees.elsevier.com/csda (Choose Article Type: Research paper, and then Select "Section IV. Annals of Statistical Data Science").

Editors: Erricos Kontoghiorghes and Ana Colubi (CMStatistics)

Guest Associate Editors: Julyan Arbel, Peter Buhlmann, Stefano Castruccio, Bertrand Clarke, Christophe Croux, Maria Brigida Ferraro, Yulia Gel Michele Guindani, Xuming He, Sangwook Kang, Ivan Kojadinovic, Chenlei Leng, Taps Maiti, Geoffrey McLachlan, Hans-Georg Mueller, Igor Pruenster, Juan Romo, Elvezio Ronchetti, Anne Ruiz-Gazen, Syvain Sardi, Xinyuan Song, Cheng Yong Tang, Roy Welsch and Peter Winker.

Contents

Saturday 14.12.2019 08:40 - 09:30 Room: Beveridge Hall Chair: Chenlei Leng Keynote talk 1

Modeling networks and network populations via graph distances

Speaker: **Sofia Olhede, EPFL, Switzerland** Simon Lunagomez, Patrick Wolfe

Networks have become a key data analysis tool. They are a simple method of characterising dependence between nodes or actors. Understanding the difference between two networks is also challenging unless they share nodes and are of the same size. We shall discuss how we may compare networks and also consider the regime where more than one network is observed. We shall also discuss how to parametrize a distribution on labelled graphs in terms of a Frechet mean graph (which depends on a user-specified choice of metric or graph distance) and a parameter that controls the concentration of this distribution about its mean. Entropy is the natural parameter for such control, varying from a point mass concentrated on the Frechet mean itself to a uniform distribution over all graphs on a given vertex set. Networks present many new statistical challenges. We shall discuss how to resolve these challenges respecting the non-Euclidean nature of network observations.

Saturday 14.12.2019 10:05 - 10:55 Room: Beveridge Hall Chair: Michael Pitt Keynote talk 2

The role of factor strength and pricing errors for estimation and inference in asset pricing models

Speaker: **M Hashem Pesaran, USC, United States** Ron Smith

The role of factor strength and pricing errors for estimation of risk premia in arbitrage asset pricing models is examined. We introduce a measure of factor strength, distinguish between observed and unobserved factors, and link the unobserved factors to the pricing errors. We show risk premia can be estimated consistently when factors are strong and pricing errors sufficiently weak, irrespective of whether returns on individual securities or portfolios are used. We then derive the distribution of two pass estimator of risk premia, allowing for non-zero pricing errors and provide an empirical application to the three Fama-French factors.

Sunday 15.12.2019 18:25 - 19:15 Room: Beveridge Hall Chair: Christopher Hans Keynote talk 3

Data-driven and science-driven Bayesian methods in astronomy and solar physics

Speaker: **David van Dyk, Imperial College London, United Kingdom**

In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Newly launched or soon-to-be launched space-based telescopes are tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogs containing terabytes of data, high resolution spectrography and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. The spectrum of new instruments is helping scientists make impressive strides in our understanding of the physical universe, but at the same time generating massive data-analytic and data-mining challenges for scientists who study the resulting data. We will illustrate and discuss the interplay of data science, machine learning, Bayesian statistics, as well as data-driven and science-driven methods in the context of several problems in astrophysics, ranging from studying the expansion history of the universe, to disentangling overlapping sources, and mapping the physical characteristics of the solar corona. A common theme involves strategies for combining multiple data sets analysed sequentially into a single coherent statistical result.

Sunday 15.12.2019 18:25 - 19:15 Room: CLO B01 Chair: Tommaso Proietti Keynote talk 4

Robust tests for white noise and cross-correlation

Speaker: **Liudas Giraitis, Queen Mary University of London, United Kingdom**

Commonly used tests to assess evidence for the absence of autocorrelation in a univariate time series or serial cross-correlation between time series rely on procedures whose validity holds for i.i.d. data. When the series are not i.i.d., the size of correlogram and cumulative Ljung-Box tests can be significantly distorted. The aim is to adapt standard correlogram and portmanteau tests to accommodate hidden dependence and non-stationarities involving heteroskedasticity, thereby uncoupling these tests from limiting assumptions that reduce their applicability in empirical work. To enhance the Ljung-Box test for non-i.i.d. data, a new cumulative test is introduced. The asymptotic size of these tests is unaffected by hidden dependence and heteroskedasticity in the series. Related extensions are provided for testing cross-correlation at various lags in bivariate time series. Tests for the i.i.d. property of a time series are also developed. An extensive Monte Carlo study confirms good performance in both size and power for the new tests. Applications to real data reveal that standard tests frequently produce spurious evidence of serial correlation.

Monday 16.12.2019 18:05 - 18:55 Room: Beveridge Hall Chair: Manfred Gilli Keynote talk 5

Text as a new source of data: First experience with conference abstracts

Speaker: **Peter Winker, University of Giessen, Germany**

The use of textual information gained momentum over the last years in economics. Text is considered as the new data in fields such as financial markets, innovation activities and economic history. For drawing meaningful conclusions from this type of data, a substantial number of steps in pre-processing and analyzing the data has to be taken. Usually, the implementation of these methods is based on previous experience, statistical methods, or human judgement. Thus, typical issues present when dealing with conventional quantitative data also apply to textual information. They might just be disguised differently, while new challenges show up. Some relevant steps in using textual data in a time series context are sketched. Abstracts of a conference series serve as an example. In particular, the following issues will be addressed: 1) selection of appropriate sources (corpora) and establishing access, 2) preparation of the textual data, 3) identification of themes, 4) quantifying the relevance of themes across documents, 5) aggregating relevance information over time. Finally, some remarks on the use of the generated indicators in further analysis will be provided. Open issues regarding, e.g. computational complexity and robustness of the methods will be discussed.

Saturday 14.12.2019 09:40 - 10:55 Parallel Session B – CMStatistics

EO172 Room CLO B01 FUNCTIONAL DATA ANALYSIS Chair: Germain Van Bever

E0624: **Prediction of missing functional data with memory**
Presenter: **Lauri Viitasaari**, Aalto University, Finland
Co-authors: Pauliina Ilmonen, Germain Van Bever, Tommi Sottinen, Nourhan Shafik
Functional observations X^i that are realisations of some Gaussian process are considered. We assume that parts of the paths are unobservable, and the aim is to fill in the missing information as accurately as possible. One natural approach is to predict some missing value X_s^k by using the information provided by $X_s^i, i \neq k$ of those functions X^i for which X_s^i is observed. However, under memory the unobserved X_s^k relies heavily on that particular observation X^k directly, and thus applying other observations X^i may be misleading, even if they are drawn from the same underlying process. We present a novel approach for accurate prediction of missing information X_s^k that is based on applying combined information provided by the observed part of the path X^k and the observed values $X_s^i, i \neq k$. Extensions beyond the Gaussianity assumption are discussed.

E0715: **Choosing among notions of statistical depth function**
Presenter: **Pavlo Mozharovskyi**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Karl Mosler
Classical statistics measures the outlyingness of a point by its Mahalanobis distance from the mean based on the covariance matrix of the data. Since the early 1990s, more general depth statistics have been developed for measuring centrality and outlyingness of multivariate data in a nonparametric way. A multivariate depth function is a function which, given a point and a distribution in d-dimensional space, yields a number between 0 and 1, while satisfying certain postulates regarding invariance, monotonicity, convexity and continuity. Accordingly, numerous notions of depth have been proposed in the literature, some of which are also robust against outlying data. The departure from classical Mahalanobis distance does not come without a cost. There is a trade-off between invariance, robustness and computational feasibility. Since recently, efficient exact and approximate algorithms for different depths and their applications have been made available in various software packages. In applications, there is a choice of a depth statistic: rather often various notions are feasible, among which we have to decide. Aspects and general principles of this choice are discussed. The speed of exact algorithms is compared. The limitations of popular approximate approaches are demonstrated, and guidelines are provided for the construction of depth-based statistical procedures, as well as for practical applications when several notions of depth appear to be computationally feasible.

E1099: **Continuous time scalar-on-function class of regression models with missing at random response**
Presenter: **Mohamed Chaouch**, United Arab Emirates University, United Arab Emirates
The focus is on nonparametric estimation of the generalized regression function based on copies of a continuous time stationary ergodic process, where the response is a missing at random real random variable, whereas the predictor takes values in some infinite-dimensional space. Pointwise and uniform consistency rates of kernel-type estimator of the regression operator are established. Asymptotic evaluations of the conditional bias and the quadratic error of the estimator are given. A Central Limit Theorem is also established. Since in practice a discretized version of the process is observed rather than a continuous one, a discussion on the sampling scheme is given, and two methods to build a confidence interval are provided. The results are stated under ergodic assumption without assuming any classical mixing conditions. Some previous results are completed and extended to the case where the predictor takes values in some infinite dimensional space and the response variable is affected by a Missing At Random mechanism.

EO645 Room MAL B02 STATISTICAL METHODS IN BIOMEDICAL STUDIES Chair: Florin Vaida

E2009: **Efficient design of longitudinal, randomized clinical trials with repeated measures**
Presenter: **Florin Vaida**, University of California San Diego, United States
The aim is to propose and analyze a novel study design, a cluster-randomized, longitudinal clinical trial with repeated measures. Units grouped within clusters are cluster randomized to two or more groups. Units are observed longitudinally, at baseline and follow-up visits. Repeated measures for the response of interest (continuous or binary variable) are obtained at each visit. We show that, counterintuitively, the best allocation schedule has fewer repeated measures at baseline than at follow-up, rather than an even allocation. The optimal degree of imbalance depends on the within-unit correlation, with the highest power gains to be made for low within-unit correlation (up to 60% improvement). Robustness of the design is also considered. Statistical analyses are based on analytical derivations and simulation.

E2011: **Imputation and post-selection inference in models with missing data**
Presenter: **Karen Messer**, University of California, San Diego, United States
It is common to encounter missing data among potential predictor variables in the setting of model selection. For example, we recently attempted to improve the commonly used risk prediction model for colorectal cancer screening, using pooled data from seven different large prospective studies. However, several important potential predictors were missing for more than half of subjects. Multiple imputation can effectively address missing data, and there are effective methods to incorporate variable selection into inference. However, there is not consensus on appropriate methods to address both issues simultaneously. We compare three approaches to such post-imputation-selection inference: a multiple- imputation approach; a single imputation-selection followed by bootstrap percentile confidence intervals; and a new bootstrap model-averaging approach. The 'Rubin's' Rules' multiple imputation estimator can have severe under coverage, and is not recommended. The imputation-selection estimator with bootstrap percentile confidence intervals works well. The bootstrap-model-averaged estimator, with the 'Efron's Rules' estimated variance, may be preferred if the true effect sizes are moderate.

E1659: **A method for testing multiple binary endpoints having a latent continuous distribution in clinical trials**
Presenter: **Takuma Ishihara**, Gifu university, Japan
In confirmatory clinical trials, two or more primary endpoints are often used to assess the efficacy of a test treatment. For example, in a clinical trial for patients with rheumatoid arthritis, a percentage of patients achieving a response of 20% improvement according to the American College of Rheumatology criteria (ACR20) in short term, and an achieving Disease Activity Score (DAS) below 3.2 in long term, are often used as primary endpoints for treatments. These endpoints can be observed as binary variables, but these variables have a latent continuous distribution. Furthermore, in general, it becomes more difficult to demonstrate that all of the endpoints are significant as the number of endpoints increases. Therefore, we propose a new test statistic for multiple binary endpoints which have a latent multivariate normal distribution in the framework. We confirm the efficacy of a test treatment when it is superior for at least one of the endpoint and not clinically inferior for the remaining endpoints. The performance of the proposed testing procedure is demonstrated through Monte Carlo simulations. We evaluate the empirical power and empirical type I error rate.

EO673 Room MAL B04 RECENT DEVELOPMENTS IN PRIVACY-PRESERVING DATA ANALYSIS Chair: Adrien Saumard

E0242: **Local differential privacy: Elbow effect in optimal density estimation and adaptation**
Presenter: **Amandine Dubois**, CREST-ENSAI, France

Co-authors: Cristina Butucea, Martin Kroll, Adrien Saumard

The problem of non-parametric density estimation is addressed under the additional constraint that only privatised data are allowed to be published and available for inference. For this purpose, we adopt a recent generalisation of classical minimax theory to the framework of local alpha-differential privacy and provide a lower bound on the rate of convergence over Besov spaces B^s_{pq} under mean integrated L_r-risk. This lower bound is deteriorated compared to the standard setup without privacy, and reveals a twofold elbow effect. A linear but non-adaptive wavelet estimator is shown to attain the lower bound whenever $p \geq r$, but provides a slower rate of convergence otherwise. An adaptive non-linear wavelet estimator with appropriately chosen smoothing parameters and thresholding is shown to attain the lower bound within a logarithmic factor for all cases.

E0979: **The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy**

Presenter: **Linjun Zhang**, Rutgers University, United States

Privacy-preserving data analysis is a rising challenge in contemporary statistics, as the privacy guarantees of statistical methods are often achieved at the expense of accuracy. We investigate the tradeoff between statistical accuracy and privacy in mean estimation and linear regression, under both the classical low-dimensional and modern high-dimensional settings. A primary focus is to establish minimax optimality for statistical estimation with the (epsilon, delta)-differential privacy constraint. To this end, we find that classical lower bound arguments fail to yield sharp results, and new technical tools are called for. We first develop a general lower bound argument for estimation problems with differential privacy constraints, and then apply the lower bound argument to mean estimation and linear regression. For these statistical problems, we also design computationally efficient algorithms that match the minimax lower bound up to a logarithmic factor. In particular, for the high-dimensional linear regression, a novel private iterative hard thresholding pursuit algorithm is proposed, based on a privately truncated version of stochastic gradient descent. The numerical performance of these algorithms is demonstrated by simulation studies and applications to real data containing sensitive information, for which privacy-preserving statistical methods are necessary.

E1510: **Estimating functionals under local differential privacy**

Presenter: **Lukas Steinberger**, University of Vienna, Austria

In this talk we discuss recent advances on a general theory for minimax optimally estimating linear and nonlinear functionals of the data generating distribution when the original data are protected by a local differential privacy constraint. We highlight how the theory of local differential privacy deviates fundamentally from the classical one and exhibit a few entirely new and perhaps surprising phenomena.

EO152 Room MAL B18 RECENT DEVELOPMENT IN BIOSTATISTICS Chair: Dongchu Sun

E1910: **Bayesian adaptive stepped wedge cluster randomized trials based on posterior predictive probability**

Presenter: **Song Zhang**, University of Texas Southwestern Medical Center, United States

Bayesian group sequential design has been widely used in clinical studies, especially in phase II and III studies. It is flexible and efficient in allowing early termination based on the accumulated data through Bayesian framework. However, so far there has been no discussion on its application to stepped wedge cluster randomized trials, which has become more popular in pragmatic trials in clinical and health care delivery studies. We propose a Bayesian strategy to design a cross-sectional stepped wedge cluster randomized trial based on posterior predictive probability. It provides additional flexibility for trialists to continuously evaluate interim observations, and make adaptive decisions accordingly to ensure trial success (for example, stopping the trial early for efficacy or futility, enrolling additional clusters, etc). Simulation algorithms and application examples are presented.

E1961: **Bayesian CUSP catastrophe model for sudden changes**

Presenter: **Zhuoqiong He**, University of Missouri, United States

The cusp catastrophe model uses a discontinuous nonlinear function to predict sudden changes. Due to the complexity of the discontinuous nonlinear relationship, there are some issues in fitting the statistical cusp regression model, i.e., gradient based optimization methods no longer work. We have developed a Bayesian method for the cusp regression model and used the posterior mean to obtain estimates of the parameters. The partial swarm optimization algorithm is used to speed up the convergence of the Markov chain Monte Carlo algorithm. The simulation study shows that the Bayesian method yields a better estimate than both the maximal likelihood estimation and the traditional stochastic differential equations method under the Maxwell convention.

E1905: **Lagged exact Bayesian online change point detection with parameter estimation**

Presenter: **Jing Cao**, Southern Methodist University, United States

Co-authors: Michael Byrd, Linh Nghiem

Identifying changes in the generative process of sequential data, known as change point detection, has become an increasingly important topic for a wide variety of fields. A recently developed approach, which we call EXact Online Bayesian Change-point Detection (EXO), has shown reasonable results with efficient computation for real time updates. The method is based on a forward recursive message-passing algorithm. However, the detected change points from these methods are unstable. We propose a new algorithm called Lagged EXact Online Bayesian Change point Detection (LEXO) that improves the accuracy and stability of the detection by incorporating l-time lags to the inference. The new algorithm adds a recursive backward step to the forward EXO and has computational complexity linear in the number of added lags. Parameter estimation is also addressed. Simulation studies with three common change point models show that the detected change points from LEXO are much more stable, and parameter estimates from LEXO have considerably lower MSE than EXO. We illustrate the applicability of the methods with two real world data examples comparing the EXO and LEXO.

EO434 Room MAL B20 STATISTICS IN PSYCHIATRY Chair: Satish Iyengar

E0271: **Covariate assisted principal regression for covariance matrix outcomes**

Presenter: **Xi Luo**, University of Texas Health Science Center at Houston, United States

Co-authors: Yi Zhao, Brian Caffo, Stewart Mostofsky

Modeling variances in data has been an important topic in many fields, including finance and neuroimaging. We consider the problem of regressing covariance matrices on vector covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. Covariate Assisted Principal (CAP) regression is introduced, which is an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the linear projections of the covariance matrices as well as the regression coefficients. We establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic resonance imaging study, our approach identifies the human brain network changes associated with age and sex.

E0298: **Mining the UK Biobank for indicators of brain health**

Presenter: **Thomas Nichols**, University of Oxford, United Kingdom

The UK Biobank contains detailed records on 1/2 million UK residents aged 40 to 69 years at baseline recruitment and, now 10 years later, MRI brain scanning on 100,000 subjects is underway. While 'only' 20,000 MRI datasets are available for our present analyses, it is still the largest such brain imaging study every undertaken. We will review several lines of work that looks to establish the link between variation in brain measures and

psychiatric outcomes and risk factors. As the coarsest measure of brain health, we examine predictors of age based solely on brain data, producing a 'brain age' estimate and examine how the difference of estimated brain age and (true) chronological age predicts psychiatric outcomes. Unlike other efforts, we pay particular attention to producing well-calibrated confidence statements on these brain age prediction. We then look at highly focused analyses to find patterns of brain structure, function and connectivity that associate with depression. While UK Biobank is a population sample with no particular focus on psychiatric disease, with so many subjects we are able to find 100's with indicators of depression and construct a matched sample. We then use multivariate methods like CCA to find brain variables that associate with depression. Overall, the UK Biobank presents a wealth of opportunities for psychiatric neuroimaging research.

E1217: **Searching for clusters that are replicable and clinically meaningful with applications to sleep health**
Presenter: **Meredith Wallace**, University of Pittsburgh, United States

Applied statisticians tasked with clustering face a sometimes-insurmountable challenge: revealing solutions that are replicable and clinically meaningful. This task becomes even more daunting as high-dimensional, multi-modal data become the norm rather than the exception, and with the realization that variables are often generated from underlying biological processes that follow skewed or non-normal distributions. This is particularly true in sleep research, where one sleeps, health can be captured across multiple dimensions and modes of measurement. Moreover, many sleep characteristics are highly skewed, even in seemingly homogeneous subgroups. We will delve into statistical challenges and potential solutions related to identifying replicable and clinically meaningful clusters in sleep health data. This will include a discussion on how to rigorously perform variable and model selection in a way that reduces spurious findings, cutting-edge methods for evaluating cluster stability and separation with non-normal distributions, establishing clinical utility of identified clusters, and the importance of replicating findings externally. We will demonstrate these challenges and potential solutions in the context of my ongoing National Institute of Aging grant, which aims to identify sleep health phenotypes and link them to prospective health outcomes in a large multi-cohort sample of community dwelling older adults.

EO671 Room MAL B35 SPATIOTEMPORAL MODELLING IN THE PRESENCE OF BIG DATA Chair: Philipp Otto

E0836: **Adaptive LASSO estimation of the spatial weights matrix**
Presenter: **Miryam Merk**, European University Viadrina, Germany

Spatial econometric research is typically based on prior knowledge of the spatial weights matrix that characterizes cross-sectional interactions. For lattice data, spatial weights are commonly determined by contiguity, nearest neighbor assumptions, economic or social characteristics. These a priori specifications may lead to misspecifications of the model parameters, which are sensitive to the choice of the spatial weights. We therefore propose to select and estimate spatial dependence structures by using an adaptive Least Absolute Shrinkage and Selection Operator (LASSO). In the case of spatio-temporal models, the spatial dependencies of the process can be identified based on their observations over time. However, for purely spatial models the number of spatial links exceeds the sample size. The spatial weights are therefore estimated by cross-sectional resampling under the identifying assumption of sparsity. The estimation procedure employs two-stage least squares (2SLS) to account for endogeneity of the spatially dependent variable.

E0948: **Spatio-temporal big data: Managing social media and trajectories**
Presenter: **Martin Werner**, Bundeswehr University Munich, Germany

The focus is first on spatial big data including organizing global datasets for getting a social media footprint for each individual building on a distributed computer based on randimized data representations and smart fault tolerance. This represents the unique challenges from the volume dimension of big data. In addition, there is a dimension of complexity, which will be shortly illustrated on trajectories by showing some methods with which quadratic and cubic settings (for example nearest neighbors) can be performed in acceptable time. Finally, we give hints on current challenges and future research including the case of quantum computing. In summary, we will discuss examples of solving big spatio-temporal data challenges through parallelization, simplification, and emerging technology.

E1164: **Modelling coastal profiles with a functional space-time model**
Presenter: **Francesco Finazzi**, University of Bergamo, Italy
Co-authors: Alessandro Fasso, Philipp Otto

Many modern environmental applications involve the collection of profile data across space and over time. In order to make spatio-temporal predictions, a statistical model able to take into account spatial and temporal correlation is needed. We discuss a functional space-time model for profile data based on latent variables with Markovian dynamics in time and spatially-correlated innovations. The model is estimated by means of the maximum likelihood approach using the expectation-maximization algorithm. Model estimation is implemented within the D-STEM software which is capable of handling complex space-time data sets with missing data. The model is applied to measured coastal profiles of the Sylt island in the German Bight between 1980 and 2017. Location and timing of the measurements are strongly related to the placement of nourishments. These are placed on a yearly basis at different locations along the island, depending on erosion losses in the previous year. Although profiles are measured on average 1-2 times per year, the spatial, temporal and functional support is heterogeneous. External forcing conditions that cause changes in the coastal profiles are measured in front of Sylt and are included in the statistical model as covariates. Furthermore, change points could be assigned to profiles that are abruptly changed by a nourishment itself.

EO144 Room MAL B36 REGULARIZATION IN ARTIFICIAL NEURAL NETWORKS Chair: Sylvain Sardy

E1700: **Conditional generation of molecules with disentangling**
Presenter: **Amina Mollaysa**, University of Geneva, University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Though machine learning approaches have shown great success in estimating properties of small molecules, the inverse problem of generating molecules with desired properties remain challenging. This difficulty is in part because the set of molecules which have a given property is structurally very diverse. Treating this inverse problem as a conditional distribution estimation task, we draw upon work in learning disentangled representations to learn a conditional distribution over molecules given a desired property, where the molecular structure is encoded in a continuous latent random variable. By including property information as an input factor independent of the structure representation, we can perform conditional molecule generation via a style transfer process, in which we explicitly set the property to a desired value at generation time. In contrast to existing approaches, we disentangle the latent factors from the property factors using a regularization term which constrains the generated molecules to have the property provided to the generation network, no matter how the latent factor changes.

E1956: **Sparsity in artificial neural networks (ANN)**
Presenter: **Sylvain Sardy**, University of Geneva, Switzerland

In the spirit of lasso, the estimation of the ANN many parameters with an L1-penalty is regularized. This has the advantage of performing variable selection (e.g., gene selection or feature selection in an image) and avoiding overfitting. The selection of the regularization parameter lambda is the Quantile Universal Threshold. This method requires no estimation of nuisance parameters (like sigma in Gaussian regression) and can retrieve the sparsity of the underlying ANN in certain regimes.

E2005: **Neural tangent kernel: Dynamics of infinitely wide DNNs**
Presenter: **Arthur Jacot**, Ecole Polytechnique Federale, Switzerland
Co-authors: Franck Gabriel, Clement Hongler

Modern deep learning has popularized the use of very large neural networks, but the theoretical tools to study such networks are still lacking. The Neural Tangent Kernel (NTK) describes how the output neutrons evolve during training. In the infinite width limit (when the number of hidden neutrons grows to infinity) the NTK converges to a deterministic and fixed limit, leading to a simple description of the dynamics of infinitely wide DNNs and showing a link to kernel methods. The NTK is affected by the architecture of the network, so it is helpful to understand how certain architecture choices affect the convergence and generalization of DNNs.

EO346 Room Bloomsbury NEW METHODOLOGIES AND ADVANCES IN SURVIVAL AND RELIABILITY Chair: Juan Eloy Ruiz-Castro

E0310: **Methods for checking the Markov condition in multi-state survival data**
Presenter: **Luis Machado**, University of Minho, Portugal

The inference in multi-state models is traditionally performed under a Markov assumption. This assumption claims that given the present state, the future evolution of the process is independent of the states previously visited and the transition times among them. Usually, this assumption is checked including covariates depending on the history. However, since previous landmark methods of the transition probabilities are free of the Markov assumption, they can also be used to introduce such tests (at least in the scope of the progressive multi-state models) by measuring their discrepancy to Markovian estimators. For each fixed value $s > 0$, local and graphical tests will be proposed in which we compare the estimated curves of the transition probabilities $p_{ij}(s,t)$, $t > s$, from the so-called Aalen-Johansen estimator (Markovian) and the landmark methods which are free of the Markov condition.

E0589: **A comparison of nonparametric bivariate survival functions under censored and truncated data**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Hongsheng Dai

Bivariate survival data have received considerable attention recently. In survival analysis it is common to deal with incomplete information of the data, due to random censoring and random truncation. Most of the existing research focuses on bivariate survival analysis when components are either censoring or truncation or when one component is censored and truncated, but the other one is fully observed. Starting from this background, we will review the most used estimators for the bivariate survival function, when both components are censored and truncated. We will explain the differences between them, focusing on their main advantages and disadvantages. Thanks to a simulation study and application to real datasets, we will compare the performance of the estimators.

E0677: **A complex multi-state redundant system with vacations in the repair**
Presenter: **Juan Eloy Ruiz-Castro**, University of Granada, Spain
Co-authors: Mohammed Dawabsha

A complex multi-state redundant system with preventive maintenance subject to multiple events is considered. The system can undergo internal failures and external shocks and several internal and external degradation levels are assumed. These are observed by random inspections, and if they are major, preventive maintenance is carried out. The repair facility is composed of a repairperson who may take one or more vacations. Two types of tasks may be performed by the repairperson, corrective repair and preventive maintenance. A policy is established for the repairpersons vacation time according to the number of units in the repair facility. All embedded times in the systems are phase-type distributed, and they are modeled by using Markov processes. The transient and stationary distribution is worked out and several interesting reliability measures are developed in a matrix-algorithmic form. Costs and rewards are included in the model. The results are implemented in Matlab. A numerical example shows the versatility of the model.

EO857 Room G11 ADVANCES IN REGRESSION DISCONTINUITY MODELS Chair: Sebastian Calonico

E0415: **Over-identified regression discontinuity design**
Presenter: **Carolina Caetano**, University of Georgia, United States
Co-authors: Gregorio Caetano, Juan Carlos Escanciano

A new identification and estimation strategy is proposed for the Regression Discontinuity Design (RDD). Our approach explores the heterogeneity in the "first stage" discontinuities for different values of a covariate to generate over-identifying restrictions. This allows us to identify quantities which cannot be identified with the standard RDD method, including the effects of multiple endogenous variables, multiple marginal effects of a multivalued endogenous variable, and heterogeneous effects conditional on covariates. Additionally, when this method is applied in the standard RDD setting (linear model with a single endogenous variable), identification relies on a weaker relevance condition and has robustness advantages to variations in the bandwidth and heterogeneous treatment effects. We propose a simple estimator, which can be readily applied using packaged software, and show its asymptotic properties. Then, we apply our approach to the problem of identifying the effects of different types of insurance coverage on health care utilization.

E0483: **Complex discontinuity designs using covariates**
Presenter: **Jose Zubizarreta**, Harvard University, United States
Co-authors: Juan Diaz

A new framework is proposed for general discontinuity designs that encompasses complex treatment rules. These rules may be determined by multiple running variables, each with many cutoffs, and that possibly lead to the same treatment. Moreover, the running variables may be discrete and the treatments do not need to be binary. In this framework, the observed covariates play a central role, and identification relies on a local unconfoundedness assumption. Estimation proceeds as in any observational study under strong ignorability, yet in a neighborhood of the cutoffs of the running variables. We discuss estimation approaches based on matching and weighting, including additional regression adjustments in doubly robust estimators. We present assumptions for generalization; that is, for identification and estimation of average treatment effects for target populations beyond the study sample that resides in a neighborhood of the cutoffs. We also examine a new approach to select the neighborhood for the analyses and assess the plausibility of the assumptions. We argue that, in a sense, traditional continuity-based and local-randomization frameworks for regression discontinuity designs are particular cases of our proposed framework. We motivate and illustrate this framework on a study of the impact of grade retention on juvenile crime.

E1213: **Selecting subpopulations for causal inference in regression-discontinuity studies**
Presenter: **Laura Forastiere**, Yale University, United States
Co-authors: Alessandra Mattei, Fabrizia Mealli

Extracting causal information from regression-discontinuity (RD) studies, where the treatment assignment rule depends on some type of cutoff formula, may be challenging, especially in the presence of big data. Following previous work, we formally describe RD designs as local randomized experiments within the potential outcome approach. Under this framework, causal inference concerns units belonging to some subpopulation where a local overlap assumption, SUTVA and a local randomization assumption (RD assumptions) hold. Unfortunately we do not usually know the subpopulations for which we can draw valid causal inference. We propose to use a model-based finite mixture approach to clustering in a Bayesian framework to classify observations into subpopulations for which we can draw valid causal inference, and subpopulations from which we can extract no causal information on the basis of the observed data and the RD assumptions. This approach has important advantages: It explicitly accounts for the uncertainty about sub-population membership; it does not impose any constraint on the shape of the subpopulation; and

it properly works in high-dimensional settings. We illustrate the framework in a high-dimensional RD study concerning the effects of the Borsa Famlia program, a social welfare program of the Brazilian government, on leprosy incidence.

EO160 Room G21A STATISTICAL AND DATA SCIENCE METHODS FOR BLOCKCHAIN DATA ANALYTICS Chair: Soohyun Choi

E1585: **The adaptive market hypothesis in the high frequency cryptocurrency market**
Presenter: **Jeffrey Chu**, Universidad Carlos III de Madrid, Spain

The adaptive market hypothesis (AMH) is investigated with respect to the high frequency markets of the two largest cryptocurrencies, Bitcoin and Ethereum, versus the Euro and US Dollar. Our findings are consistent with the AMH, and show that the efficiency of the markets varies over time. We also discuss possible news and events which coincide with significant changes in the market efficiency. Furthermore, we analyse the effect of the sentiment of these news and other factors (events) on the market efficiency in the high frequency setting, and provide a simple event analysis to investigate whether specific factors affect the market efficiency/inefficiency. The results show that the sentiment and types of news and events may not be a significant factor in determining the efficiency of cryptocurrency markets.

E1644: **Forecasting Ethereum STORJ token prices: Comparative analyses of applied bitcoin models**
Presenter: **Rhonda Bush**, University of Texas at Dallas, United States
Co-authors: Soohyun Choi

The research on forecasting Ethereum STORJ token has not been widely studied compared to forecasting Bitcoin. We evaluate the utility and limitations of existing Bitcoin price forecasting models applied to the Ethereum STORJ token price. We evaluate the dynamics of the model predictive utility across three time horizons ($h = 5$ days, h= 20days and $h = 50$ days), and we determine if Ethereum STORJ token clustering coefficients impacted the effectiveness of the forecasting models.

E1745: **Predicting the price of Bitcoin by its transaction network**
Presenter: **Marcell Tamas Kurbucz**, University of Pannonia, Hungary

Studies on the Bitcoin transaction network have increased rapidly in recent years, but still, little is known about the networks influence on Bitcoin prices. The goals are twofold: to determine the predictive power of the transaction networks most frequent edges on the future price of Bitcoin and to provide an efficient technique for applying this untapped dataset in day trading. A complex method consisting of single-hidden layer feedforward neural networks (SLFNs) is used. The presented method achieved an accuracy of approximately 60.05% during daily price movement classifications, despite only considering a small subset of edges.

EO498 Room G3 MODELING COMPLEX DATA STRUCTURE WITH APPLICATIONS Chair: Yunpeng Zhao

E0304: **On the progress to reduce economic inequality: Insights from recently proposed measures of inequality**
Presenter: **Joseph Gastwirth**, George Washington University, United States

Most measures of income inequality compare a measure of the variation in income to the average. When most of the increase in income goes to the upper part of the curve, this leads to an under-estimate of the true change. One new measure, G2, replaces the mean by the median in the Gini index and shows that income inequality in the U.S., U.K and other nations increased faster than the Gini index. Two transformations of the Lorenz curve asks what fraction of income cumulated from the bottom or middle is needed to have the same share of income as the top 100q%. These curves and related area measures also indicate a greater rate of increase in inequality. Recent extensions of these curves can be used to examine the progress of minorities and women. They show that in the U.S., since 2000, African-Americans have made very little progress relative to whites. While the income of women has increased somewhat relative to males, most of these relative gains were received by women in the upper portion of the female distribution.

E0491: **CESME: Cluster analysis with latent semiparametric mixture models**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Hui Zou, Lyuou Zhang, Lulu Wang

Model-based clustering is one of the most popular statistical approaches in unsupervised learning and has been widely employed in practice for exploratory analysis, data visualization, sub-community identification, and quality control. Regardless of its wide applicability, the traditional distributional assumption of Gaussianity is too stringent to be validated in general, and therefore prevents the model-based clustering to be used for data with complex distributions, such as high skewness. We propose a flexible semiparametric latent model to cluster multivariate data deviated from Gaussian. The model assumes that the observed random vectors are obtained from unknown monotone transformations of latent variables governed by a Gaussian mixture distribution. The identifiability of the proposed model is carefully studied. An alternating maximization procedure is developed to estimate the proposed model, whose convergence property is investigated by using finite-sample analysis. An interesting transition phenomenon of the convergence for the proposed algorithm, which is due to the presence of the unknown transformations, is explored and provides guidance on the design of the algorithm. The proposed method is also numerically assessed through extensive simulations, and demonstrates superior performance compared to most of the contemporary competitors.

E1815: **Graph theory and combinatorics for group-level network inference**
Presenter: **Shuo Chen**, University of Maryland, School of Medicine, United States

The focus is on group-level statistical inference for networks, where outcomes are multivariate edge variables constrained in an adjacency matrix. The graph notation is used to represent the network outcome variables, where nodes are identical biological units (e.g. brain regions) shared across subjects and edge-variables indicate the strengths of the interactive relationships between nodes. The edge-variables vary across subjects and may be associated with covariates of interest. The statistical inference for multivariate edge-variables is challenging because both localized inference on individual edges and the joint inference of a combinatorial of edges (network-level) are desired. We develop a group-level network inference model to integrate graph theory and combinatorics into group-level network statistical inference. We first propose an objective function with 0 norm regularization to capture latent subgraphs/subnetworks accurately by suppressing false positive edges. We next statistically test each detected subnetwork using graph combinatorics based statistical inferential procedure. The results demonstrate the proposed method outperform existing multivariate statistical methods by simultaneously reducing false positive and false negative discovery rates and increasing replicability.

EO596 Room G4 PERMUTATION TESTS Chair: Olivier Renaud

E0932: **Robust testing in generalized linear models by sign-flipping score contributions**
Presenter: **Livio Finos**, University of Padova, Italy

Generalized linear models are often misspecified due to overdispersion, heteroscedasticity and ignored nuisance variables. Existing quasi-likelihood methods for testing in misspecified models often do not provide satisfactory type-I error rate control. We present a novel semi-parametric test, based on sign-flipping individual score contributions that is proven to be robust against variance misspecification. When nuisance parameters are estimated, our basic test becomes conservative. We show how to take nuisance estimation into account to obtain an asymptotically exact test. The speed of convergence can be further accelerated considering a particular transformation of the contributions of the score that makes them independent. With this transformation, the test shows an excellent control of the first type error, even for very low sample size and nuisances parameters strongly correlated with the tested parameter. The advantage with respect to other methods is further magnified when the method is

extended to the multi-dimensional and even high-dimensional setting. The presented approach is very flexible. We show natural extensions to more complex models, such as random effect models and penalized models.

E0971: **Permutation-based prediction bands for functional data: A conformal prediction approach**
Presenter: **Simone Vantini**, Politecnico di Milano, Italy
Co-authors: Matteo Fontana, Alexander Gammerman, Vladimir Vovk

The focus will be on the prediction of a new unobserved functional datum given a set of observed functional data, possibly in presence of covariates, either scalar, categorical, or functional. In particular, we will present an approach (i) able to provide prediction regions which could visualized in the form of bands, (ii) guaranteed with exact coverage probability, (iii) not relying on parametric assumptions about the specific distribution of the functional data set, and finally (iv) being computational efficient. The method is built on a combination of ideas coming from the recent literature pertaining to functional data analysis (i.e., the statistical analysis of datasets made of functions) and conformal prediction (i.e., a predictive version of permutation tests). We will present some simulations enlightening the flexibility of the approach and the effect on the amplitude of prediction bands of different algorithmic choices. Finally, we will apply the method to some benchmark case studies and to a more thorough application pertaining to the prediction of time varying mobility flows within the city of Milan.

E1194: **Testing time by time differences of EEG signals using the slopes within multiple comparisons procedures**
Presenter: **Jaromil Frossard**, University of Geneva, Switzerland
Co-authors: Olivier Renaud

Electroencephalography (EEG) is widely used in experiments recording brain activity. The researcher would like to compare the EEG signals and to test differences between experimental conditions. Each time point is tested, and a multiple comparisons procedure should be used to control the family wise error rate (FWER). The state-of-the-art statistical analysis is permutation tests in combination with multiple comparisons procedures like the cluster-mass tests or the threshold-free cluster-enhancement. We show how the slopes of the signals may be used to better modelise the researcher's hypothesis and to increase the power of the test while controlling the FWER.

EO446 Room Gordon ODE INFERENCE AND APPLICATIONS Chair: Nicolas Brunel

E0495: **Deep learning as optimal control: Models and numerical methods**
Presenter: **Elena Celledoni**, Norwegian University of Science and Technology, Norway

Deep learning neural networks have been recently interpreted as discretisations of an optimal control problem subject to an ordinary differential equation constraint. We review the first order conditions for optimality, and the conditions ensuring optimality after discretization. This leads to a class of algorithms for solving the discrete optimal control problem which guarantee that the corresponding discrete necessary conditions for optimality are fulfilled. We discuss two different deep learning algorithms and make a preliminary analysis of the ability of the algorithms to generalise.

E0524: **Parameter estimation in mixed effect models based on ordinary differential equations: An optimal control approach**
Presenter: **Quentin Clairon**, ISPED - Universita de Bordeaux, France
Co-authors: Chloe Pasin, Melanie Prague, Irene Balelli, Rodolphe Thiebaut

A parameter estimation method is presented for nonlinear mixed effect-models based on ordinary differential equations (NLME-ODEs). These models aim to describe the dynamic of a whole population while accounting for the observed variability between subjects. Their relevance for the analysis of biological processes implying a large number of subjects but limited individual measurements during time has already led to the development of parameter estimation methods based on stochastic algorithms. However, these methods generally: 1) do not consider potential model misspecifications; 2) need to estimate initial conditions for each patient or make strong assumptions on their values; 3) show dramatic degradation of their accuracy in presence of poorly identifiable parameters. To face these problems, we propose an original method based on discrete optimal control theory. This procedure incorporates a possible gap between the model describing the population dynamic and the specific individual dynamic. In addition, it is based on a profiled cost function on the initial conditions to avoid their estimation. We compare our approach with other ones on a model proposed to study the antibody concentration dynamics after vaccination against Ebola virus.

E1004: **The Frenet-Serret equation for the estimation of the mean shape of multidimensional functional data**
Presenter: **Nicolas Brunel**, ENSIIE, France
Co-authors: Juhyun Park

Variations of univariate curves can be efficiently analysed with functional data analysis tools, such as functional PCA or square root velocity transform to name a few. When considering multidimensional curves, these statistical methods can be generalised in several directions, as the notion of multidimensional amplitude can be defined in several ways. We propose a geometrical framework for the analysis of multidimensional functional data, where we aim at estimating the variations of shapes. As a first step, we define a mean shape for a population of curves. For space curves, we need to estimate curvatures and torsions, which are known to be very challenging when using discrete and noisy observations, as their definition involves higher order derivatives. We use the Frenet-Serret formula that defines an Ordinary Differential Equation in the Lie group or rotations, and we show that estimating the geometry of curve is equivalent to estimate time-varying coefficients. We build a penalised criterion and an adaptive estimator of the curvature, torsion and shape and derive an iterative algorithm based on the Magnus expansion. Finally, different simulation settings and real data case are discussed.

EO552 Room MAL G14 ADVANCES IN MIXED MODELS Chair: Garth Tarr

E0555: **Bootstrapping clustered data via a weighted Laplace approximation**
Presenter: **Daniel Antonio Flores Agreda**, Universite de Geneve, Switzerland
Co-authors: Eva Cantoni, Stephane Heritier, Rory Wolfe

The problem of bootstrapping Generalized Linear Mixed Models for exponential families is considered in a non-parametric manner. We propose a method based on the random weighting of the individual contributions to the joint distribution of outcomes and random effects and the use of the Laplace approximation method for integrals on this weighted joint distribution. We show the similarities between the random weighting of the Laplace-approximated log-Likelihood and other bootstrap schemes based on random weighting of the estimating equations. Through simulations, we provide evidence of the good quality of the bootstrap approximations of the sampling distributions for the model parameters as well as evidence of their finite sample properties when applied in a Mixed Logit Model. We further illustrate the properties of our proposal via simulated examples in Accelerated Failure Time Models for clustered data.

E0648: **Conditional model selection in mixed-effects models with cAIC4**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Benjamin Saefken, David Ruegamer, Thomas Kneib

Model selection in mixed models based on the conditional distribution is appropriate for many practical applications and has been a focus of recent statistical research. We introduce the R package cAIC4 that allows for the computation of the conditional Akaike Information Criterion (cAIC). Computation of the conditional AIC needs to take into account the uncertainty of the random effects variance and is therefore not straightforward. We introduce a fast and stable implementation for the calculation of the cAIC for (generalized) linear mixed models estimated with lme4 and

(generalized) additive mixed models estimated with gamm4. Furthermore, cAIC4 offers a stepwise function that allows for an automated stepwise selection scheme for mixed models based on the cAIC. Examples of many possible applications are presented to illustrate the practical impact and easy handling of the package.

E0896: **Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models**
Presenter: **Sahir Bhatnagar**, McGill University, Canada
Co-authors: Karim Oualkacha, Yi Yang, Tianyuan Lu, Erwin Schurr, JC Loredo-Osti, Marie Forest, Celia Greenwood

Complex traits are known to be influenced by a combination of environmental factors and rare and common genetic variants. However, detection of such multivariate associations can be compromised by low statistical power and confounded by population structure. Linear mixed effect models (LMM) can account for correlations due to relatedness, but have not been applicable in high-dimensional (HD) settings where the number of fixed effect predictors greatly exceeds the number of samples. False positives can result from two-stage approaches, where the residuals estimated from a null model adjusted for the subjects' relationship structure are subsequently used as the response in a standard penalized regression model. To overcome these challenges, we develop a general penalized LMM framework called ggmix for simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. We develop a blockwise coordinate descent algorithm which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Through simulations and two real data examples, we show that ggmix leads to better sensitivity and specificity compared to the two-stage approach or principal component adjustment while maintaining good predictive ability. ggmix can be used to construct polygenic risk scores and select instrumental variables in Mendelian randomization studies.

EO558 Room MAL G15 BAYESIAN MACHINE LEARNING Chair: Julyan Arbel

E1592: **Combining model and parameter uncertainty in Bayesian neural networks**
Presenter: **Geir Olve Storvik**, University of Oslo, Norway
Co-authors: Aliaksandr Hubin

Bayesian neural networks (BNNs) have recently regained a significant amount of attention in the deep learning community due to the development of scalable approximate Bayesian inference techniques. There are several advantages of using a Bayesian approach: Parameter and prediction uncertainties become easily available, facilitating rigid statistical analysis. Furthermore, prior knowledge can be incorporated. However, so far, there have been no scalable techniques capable of combining both model (structural) and parameter uncertainty. We introduce the concept of model uncertainty in BNNs, and hence we make inference in the joint space of models and parameters. Moreover, we suggest an adaptation of a scalable variational inference approach with reparametrization of marginal inclusion probabilities to incorporate the model space constraints. Finally, we show that incorporating model uncertainty via Bayesian model averaging and Bayesian model selection allows us to drastically sparsify the structure of BNNs.

E1984: **Partially exchangeable networks and architectures for learning summary statistics in approximate Bayesian computation**
Presenter: **Pierre-Alexandre Mattei**, INRIA, Universite Cote d'Azur, France
Co-authors: Samuel Wiqvist, Umberto Picchini, Jes Frellsen

A novel family of deep neural architectures, named Partially Exchangeable Networks (PENs) that leverage probabilistic symmetries, is presented. By design, PENs are invariant to block-switch transformations, which characterize the partial exchangeability properties of conditionally Markovian processes. Moreover, we show that any block-switch invariant function has a PEN-like representation. The DeepSets architecture is a special case of PEN and we can therefore also target fully exchangeable data. We employ PENs to learn summary statistics in approximate Bayesian computation (ABC). When comparing PENs to previous deep learning methods for learning summary statistics, our results are highly competitive, both considering time series and static models. Indeed, PENs provide more reliable posterior samples even when using less training data.

E1989: **Interpreting deep neural networks through variable importance**
Presenter: **Jonathan Ish-Horowicz**, Imperial College London, United Kingdom
Co-authors: Seth Flaxman, Sarah Filippi

While the success of deep neural networks is well-established across a variety of domains, the ability to explain and interpret these methods is limited. Unlike previously proposed local methods which try to explain particular classification decisions, we focus on global interpretability and ask a universally applicable, and surprisingly understudied question: given a trained model, which features are the most important? In the context of neural networks, a feature is rarely important on its own, so our strategy is specifically designed to leverage partial covariance structures and incorporate variable interactions into our proposed feature ranking. The methodological contributions are three-fold. First, we propose a novel effect size analogue for the problem of global interpretability, which is appropriate for applications with highly collinear predictors (ubiquitous in computer vision). Second, we extend the recently proposed "RelATive cEntrality" (RATE) measure to the Bayesian deep learning setting. RATE applies an information theoretic criterion to the posterior distribution of effect sizes to assess feature significance. Unlike competing methods, our method has no tuning parameters to pick or costly randomization steps. Overall, we show state-of-the-art results applying our framework to several application areas including: computer vision, genetics, natural language processing, and social science.

EO779 Room MAL G16 THE TIME ISSUE FOR COMPLEX DATA FROM HUMANITIES AND SOCIAL SCIENCES Chair: Madalina Olteanu

E1506: **Change detection and clustering in temporal graphs**
Presenter: **Fabrice Rossi**, Universite Paris Dauphine, France
Co-authors: Pierre Latouche, Marco Corneli

Graphs are commonly used to represent interactions between entities, interactions which can be repeated and happen at specific times. This leads naturally to the concept of temporal graphs. In general, those graphs are represented as a time series of static graphs using a crude time quantization technique: the data analyst chooses a timescale and disregards temporal information at a finer scale. For instance, one can produce daily interaction graphs. While this approach can produce interesting results, it cannot adapt to more complex schemes where a single time scale cannot capture the full temporal dynamic of the interactions. An alternative generative model is described that models directly the temporal structure of the dynamic graph. The model is based on the principle of the stochastic block model and extends the static setting to a temporal one by describing interactions between two classes of vertices via a non homogeneous Poisson point process (NHPPP). The complexity of those NHPPPs will be controlled by enforcing a piecewise constant structure on the intensity functions, with globally shared intervals. As a consequence, the estimation of the model will provide both a clustering structure for the vertices and time intervals in which all the NHPPPs will have constant but distinct intensities. This latter structure could be used to produce a time series of graphs with stationary interaction structure, leading to an automated local time scale analysis.

E1536: **Markov and the Dukes of Savoy: A temporal analysis of the Piedmontese-Savoyard legislation**
Presenter: **Madalina Olteanu**, Pantheon-Sorbonne University, France
Co-authors: Julien Alerini

Although time is at the core of historian's research, using the right measure or the right scale for studying it are still problematical. One common approach consists in an a priori chronicle deconstruction, extracting the high points which are sometimes perceived as changeover instants between historical periods. When one is studying a time series, the employed approach and results are always "controlled" by a general context, and one usually looks for the relation between the context and the chronicle. The time series quantifying the legislation and more particularly the legislation

related to military logistics, issued by the Duchy of Savoy during the XVIth and the XVIIth centuries, is not an exception in this regard, and is generally read according to the vicissitudes of peace and war. We aim at illustrating how hidden Markov models and integer-valued autoregressive models provide new ways and perspectives for establishing new periodizations and temporalities of the State, hinging on the transition phases identified by the models.

E1537: **Detecting the evolution phases of a text production**
Presenter: **Stephane Lamasse**, Universita Paris 1, France
Co-authors: Madalina Olteanu

The aim is to illustrate how one may identify the transformations of a text over time. We investigate the content of the Wikipedia pages of several famous researchers and historical figures in order to bring out their production phases. Temporal information on the content of a page, since its creation and with a high temporal resolution, may be available: the size of the page, the number of words, and, with more text-mining effort, the table of contents, ... We apply time-segmentation techniques (change-point detection) and semantic analysis for exploring the evolution of the pages and identifying key-events.

EO781 Room Woburn INFERENCE FOR IMPRECISE AND INDIRECT DATA Chair: Taoufik Bouezmarni

E1607: **Semiparametric estimation of the distribution of episodically consumed foods measured with error**
Presenter: **Felix Camirand Lemyre**, Universita de Sherbrooke, Canada
Co-authors: Raymond Carroll, Aurore Delaigle

One purpose of collecting dietary data in national surveys has been to estimate the distribution of usual intake of various nutrients and food groups in populations and subpopulations. When dietary intake data are obtained by 24-hour dietary recall, it is now widely recognized that the usual intake is assessed with considerable measurement error. There is a vast literature on measurement errors for estimating usual intake distributions when the food/nutrient is consumed daily. However, the classical methods cannot be used when considering the usual intake of episodically consumed foods (e.g., fish or whole fruits) because in that case those methods are typically biased. We address this problem by using a mixture of a parametric and a non-parametric approach.

E1626: **From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data**
Presenter: **Christopher McMahan**, Clemson University, United States
Co-authors: Joshua Tebbs, Christopher Bilder, Chase Joyner

Due to reductions in both time and cost, group testing is a popular alternative to individual-level testing for disease screening. These reductions are obtained by testing pooled biospecimens (e.g., blood, urine, saliva, etc.) for the presence of an infectious agent. However, these reductions come at the expense of data complexity, making the task of conducting disease surveillance more tenuous when compared to using individual-level data. This is because an individual's disease status may be obscured by a group testing protocol and the effect of imperfect testing. Further, unlike individual-level testing, a given participant could be involved in multiple testing outcomes and/or may never be tested individually. To circumvent these complexities and to incorporate all available data, we propose a Bayesian generalized linear mixed model that accommodates data arising from any group testing protocol, estimates unknown assay accuracies, and accounts for the potential heterogeneity in the covariate effects across population subgroups (e.g., clinic sites); this latter feature being of key interest to practitioners tasked with conducting disease surveillance. To achieve model selection, our proposal uses spike and slab priors for both fixed and random effects. The developed methodology is illustrated through numerical studies and is applied to chlamydia surveillance data collected in Iowa.

E1673: **A test of exogeneity in the functional linear regression model**
Presenter: **Melanie Birke**, University of Bayreuth, Germany
Co-authors: Manuela Dorn, Carsten Jentsch

Models containing endogenous control variables often occur in econometrics, natural sciences and other disciplines. They usually require more complex estimation methods. If the endogeneity remains unnoticed, it may lead to inconsistent estimates. In multivariate statistics several methods for testing exogeneity are known yet. We, in contrast, focus on the functional linear regression model, where the slope parameter belongs to the Sobolev space of periodic functions. For the functional linear model there exist estimation methods for the exogenous as well as the endogenous case, but no test for exogeneity is known so far. Assuming that an optimal linear instrument for the endogenous control variable exists, there is literature about instrumental variable estimators which are consistent in the endogenous case, whereas the ordinary least-squares estimator also proposed in literature for the exogenous case, is inconsistent under endogeneity. Based on the idea of the Hausman test, we compare both estimators to introduce a test for exogeneity. However, some modifications on the test statistic are necessary, since a direct analogue of the one used in the original Hausman test is not applicable in the functional context. We show the asymptotic behavior of the test statistic as well as the consistency of the bootstrap analogue. Finally, the test's finite-sample performance is checked by a small simulation study.

EO566 Room Chancellor's Hall SPATIAL STATISTICS Chair: Mattias Villani

E0672: **Determinantal point processes and their usefulness in spatial statistics**
Presenter: **Jesper Moeller**, Aalborg University, Denmark

The purpose is to discuss how the appealing properties of determinantal point processes (DPPs) can be used for constructing statistical models and methods, exploiting that the likelihood and moment expressions can be evaluated and realizations can be simulated in a simple way where freely available software have been developed for DPPs in space or on the sphere. We pay particular attention to DPPs in space, and we also discuss to what extent DPPs are useful for the description of repulsiveness.

E0945: **Gaussian variational approximations for high-dimensional state space models**
Presenter: **Matias Quiroz**, University of Technology Sydney, Australia
Co-authors: David Nott, Robert Kohn

Variational approximations of the posterior distribution in high-dimensional state space models, which encompass spatio-temporal models, are considered. The variational approximation is a multivariate Gaussian density, in which the variational parameters to be optimized are a mean vector and a covariance matrix. The number of parameters in the covariance matrix grows as the square of the number of model parameters, so it is necessary to find simple yet effective parameterizations of the covariance structure when the number of model parameters is large. The joint posterior distribution over the high-dimensional state vectors is approximated by a dynamic factor model, with Markovian time dependence and a factor covariance structure for the states. This gives a reduced dimension description of the dependence structure for the states, as well as a temporal conditional independence structure similar to that in the true posterior. We consider an application that models the spread of the Eurasian Collared-Dove across North America.

E1185: **The rational SPDE approach for Gaussian random fields with general smoothness**
Presenter: **David Bolin**, University of Gothenburg, Sweden
Co-authors: Kristin Kirchner

A popular approach for modeling and inference in spatial statistics is to represent Gaussian random fields as solutions to stochastic partial differential equations (SPDEs) of the form $L^\beta u = \mathcal{W}$, where $\mathcal{W}$ is Gaussian white noise, L is a second-order differential operator, and $\beta > 0$ is a parameter

that determines the smoothness of u. However, this approach has been limited to the case $2\beta \in \mathbb{N}$, which excludes several important models and makes it necessary to keep β fixed during inference. We propose a new method, the rational SPDE approach, which in spatial dimension $d \in \mathbb{N}$ is applicable for any $\beta > d/4$, and thus remedies the mentioned limitation. The presented scheme combines a finite element discretization with a rational approximation of the function $x^{-\beta}$ to approximate u. For the resulting approximation, an explicit rate of convergence to u in mean-square sense is derived. Furthermore, we show that our method has the same computational benefits as in the restricted case $2\beta \in \mathbb{N}$. Numerical experiments and a statistical application are used to illustrate the accuracy of the method, and to show that it facilitates likelihood-based inference for all model parameters including β.

EO102 Room CLO 101 HIGH DIMENSIONAL PROBLEMS WITH BIOLOGICAL APPLICATIONS Chair: Juhyun Park

E0573: **A methodology to select and rank covariates in high-dimensional data under dependence**
Presenter: **Anne Gegout-Petit**, INRIA Universite de Lorraine-IECL BIGS, France
Co-authors: Aurelie Muller-Guedin

A methodology is proposed to select and rank covariates associated with a variable of interest in a context of high-dimensional data under dependence but few observations. The methodology imbricates successively clustering of covariates, decorrelation of covariates using Factor Latent Analysis, selection using aggregation of adapted methods and finally ranking. We present "armada", the package associated with the method. The performance of the method is assessed by simulations. An application to a real case in the framework of personalised medicine is given: the purpose is to find omics data linked with a biomarker of breast cancer.

E1201: **Spatially correlated functional data analysis: Application to brain imaging**
Presenter: **Surajit Ray**, University of Glasgow, United Kingdom
Co-authors: Salihah Alghamdi

A model is proposed for analysing replicated functional data which are spatially correlated. The research stems from the need for accurate estimation of spatio-temporal fields by summarising information observed over several replicates. The framework generalizes the existing framework of spatio-temporal regression model with partial differential equations regularisation (ST-PDE) approach, and thus can accommodate spatially dependent functions or time dependent surfaces embedded in manifolds and irregular boundaries. This need has emerged from a study on classification of brain signals based on the difference in visual stimulus. Analytically, we show that the estimators of composite spatio-temporal field is relatively more efficient than existing estimators. The proposed method is thoroughly compared via simulation studies to existing spatio-temporal functional techniques and is applied to the analysis of the EEG data on brain signals to provide a composite temporally varying brain map over several replications.

E1038: **Multilevel multiclass Gaussian graphical model**
Presenter: **Inyoung Kim**, Virginia Tech, United States

Gaussian graphical models have been a popular tool to investigate conditional dependency structure between random variables by estimating sparse precision matrices. The estimated precision matrices could be mapped into networks for visualization. However, investigating the conditional dependency structure when there exists the two-level structure among variables is still limited; some variables are considered as higher level variables while others are nested in these higher-level variables; the latter are called lower-level variables. For instance, genes are grouped into pathways for particular functions, so that pathways are the higher-level variables and genes within pathways are the lower level variables. Higher-level variables are not isolated; instead, they work together to accomplish certain tasks. Therefore, simultaneously exploring conditional dependency structures among higher level variables and among lower level variables are of our main interest. Given two level data from heterogeneous classes, we propose a method to jointly estimate the two level Gaussian graphical models across multiple classes, so that common structures in terms of the two level conditional dependency are shared during the estimation procedure, yet unique structures for each class are retained as well. Our proposed approach is achieved by first introducing higher-level variable factors within classes, and then introducing common factors across classes.

EO194 Room CLO 102 SCALABLE STATISTICAL METHODS Chair: Cristiano Varin

E1274: **Statistical scalability of approximate inference for spatial models**
Presenter: **Helen Ogden**, University of Southampton, United Kingdom

Even for simple normal models for spatial data, calculating the likelihood function can be infeasible for large datasets, as the cost of calculating the likelihood grows cubically with the number of data points. Because of this, many approximations to the likelihood have been proposed, all designed to be computationally scalable, so that the cost of computing the likelihood approximation does not grow too quickly with the size of the data. We study the statistical properties of inference with several families of approximate likelihoods, each of which involves tuning parameters which control a trade-off between computational cost and accuracy of the approximation. We discuss how the tuning parameters should be chosen to maintain good statistical properties as the amount of data grows, and discuss the implications of this for the scalability of the resulting inference.

E1787: **Scalable and adaptive smooth modelling**
Presenter: **Alkeos Tsokos**, UCL, United Kingdom
Co-authors: Ioannis Kosmidis

An approach is developed to modelling with smooth components that produces locally adaptive fits while being tuning parameter free. The approach relies on representing smooth functions as linear combinations of b-splines whose coefficients are equipped with a particular sparsity inducing prior. Because the approach is tuning parameter free, it scales well with the number of smooth components estimated. We provide an efficient algorithm to compute the estimates and apply the method to additive modelling and scalar on function regression. We demonstrate its performance on simulated data and show some examples of real world use.

E1996: **Pairwise likelihood methods for paired comparison models**
Presenter: **Manuela Cattelan**, University of Padova, Italy

Paired comparisons data are binary or categorical data representing the results of comparisons of a set of items performed two by two. Usually, observations are dependent either because the same person performs multiple paired comparison, or because the same item is involved in more than one comparison. A typical example of the latter instance is a round robin tournament in which each player is involved in different competitions, one against each other player. Hence, models that account for dependence in paired comparison data should specify a multivariate distribution for binary or categorical data whose dimension may be very large. Inferential procedures for this type of models become quickly cumbersome as the number of items involved in the paired comparisons increases. The aim is to present how pairwise likelihood methods can be employed to overcome inferential difficulties in such models. Moreover, the need to specify only bivariate distributions reduces the problems related to the growth of the number of items involved. The proposed methodology will be illustrated through applications to real data sets.

EO841 Room Court MODELLING AND CLUSTERING COMPLEX DATA I Chair: Geoffrey McLachlan

E1342: **Mixture of autoregressive moving average models**
Presenter: **Hien Nguyen**, La Trobe University, Australia

In real-world modelling, time series are often heterogeneous and may take on the features of different regimes over different points in time. Modelling these possible different regimes can be done using a mixture variant of the traditional method of autoregressive moving average modelling. In recent times, autoregressive variants have been proposed due to the ease of estimation of such models using maximum composite likelihood estimation EM procedure. However, the moving average elements cannot be so easily estimated in in this way. We demonstrate how new advances in automatic differentiation can be used to conduct such estimation.

E0774: **bHUB: Bayesian inference of hub nodes across multiple networks with application to ovarian cancer**
Presenter: **Kim-Anh Do**, University of Texas MD Anderson Cancer Center, United States
Co-authors: Min Jin Ha, Christine Peterson

Hub nodes within biological networks play a pivotal role in determining phenotypes and disease outcomes. In the multiple network setting, we are interested in understanding network similarities and differences across different experimental conditions or subtypes of disease. The majority of proposed approaches for clustering and joint modeling of multiple networks focus on the sharing of edges across graphs. Rather than assuming the network similarities are driven by individual edges, we instead focus on the presence of common hub nodes, which are more likely to be preserved across settings. Specifically, we formulate a Bayesian approach to the problem of multiple network inference which allows direct inference on shared and differential hub nodes. The proposed method not only allows a more intuitive interpretation of the resulting networks and clearer guidance on potential targets for treatment, but also improves power for identifying the edges of highly connected nodes. Through simulations, we demonstrate the utility of our method and compare its performance to current popular methods that do not borrow information regarding hub nodes across networks. We illustrate the applicability of our method to inference of co-expression networks from The Cancer Genome Atlas ovarian carcinoma data set.

E1379: **Modelling and clustering of private distributed data**
Presenter: **Sharon Lee**, University of Adelaide, Australia

As collaborative data analysis is increasingly common, privacy has become a major concern. However, performing modelling and clustering on distributed data without disclosing the full data is a challenging task, especially in the commercial and healthcare settings where strict privacy agreements and policies must be followed. We present a privacy-enhanced modification of the EM algorithm for fitting mixture models in a multi-party setting. In particular, the scenario of horizontally partitioned data is addressed. Building on the concept of secure sum computation and adopting a cyclic communication network, the proposed two-cycle M-step approach offers protection against information leakage in the case of corrupted parties. The effectiveness of our methodology will be demonstrated through a comparative security analysis, illustrated using the normal and t-mixture models on real data.

EO190 Room Jessel NEW METHODS FOR DEPENDENT DATA MODELING Chair: Cheng Yong Tang

E0221: **Model-based clustering of high-dimensional longitudinal data**
Presenter: **Luoying Yang**, University of Rochester, United States
Co-authors: Tongtong Wu

A model-based clustering method is introduced with model and variable selection for high-dimensional longitudinal data. The motivation comes from the Trial of Activity in Adolescent Girls (TAAG), which aimed to examine multi-level factors related to the change of physical activities by following up a cohort of 783 girls over 10 years from adolescence to early adulthood. The goal is to identify the intrinsic grouping of subjects with similar patterns of physical activity trajectories and the most relevant predictors among over 800 candidate variables within groups. The previous analyses could only allow clustering and variable selection conducted over two steps, while this method can perform the tasks simultaneously. By assuming each subject is drawn from a finite Gaussian mixture distribution, model effects and cluster labels are estimated based on the restricted maximum log-likelihood, with SCAD penalty and group lasso penalty applied on the fixed effects and random effects, respectively, to induce sparsity in predictors for efficient parameter estimation and identification. Bayesian Information Criterion is used to determine the optimal cluster number and tuning parameters values for the penalties. Our numerical studies show that the new model has advantages such as faster computation and more accurate clustering over other existing clustering methods and is able to accommodate complex data with multi-level and longitudinal effects.

E0813: **Information criteria for latent factor models: A study with general factor pervasiveness and adaptivity**
Presenter: **Xiao Guo**, University of Science and Technology of China, China
Co-authors: Cheng Yong Tang

The purpose is to study the information criteria for latent factor models, allowing large number of variables diverging with the sample size. Without requiring the factor pervasiveness condition as in existing studies, the proposal accommodates generic schemes that broadly and flexibly characterize the contributions from the latent common factors. We dedicatedly analyze the properties of the information criteria, and provide new insights on the fundamental importance from adequately incorporating the impact due to different strength of the factor pervasiveness. Based on the analysis, we then propose a class of new information criteria, adaptive to the strength of the factor pervasiveness, for identifying the number of the latent common factors. The theory establishes the consistency of the proposed adaptive information criteria in correctly determining the number of the latent factors. The analysis reveals that the adaptivity to the factor pervasiveness is indeed a key for the information criteria to be consistent in broad settings. As another new discovery, we show that when the strength of the factor pervasiveness is weak below certain level, then correctly determining the number of the latent factors is not feasible with information criteria. We demonstrate the performance of the new adaptive information criteria with extensive numerical examples including simulations and real data analysis.

E1399: **A predictive time-to-event modeling approach with longitudinal measurements and missing data**
Presenter: **Cheng Yong Tang**, Temple University, United States

An important practical problem in survival analysis is predicting the time to a future event such as the death or failure of a subject. It is of great importance for medical decision making to investigate how the predictor variables including repeated measurements of the same subjects are affecting future time-to-event. Such a prediction problem is particularly more challenging due to the fact that the future values of predictor variables are unknown, and they may vary dynamically over time. We consider a predictive approach based on modeling the forward intensity function. To handle the practical difficulty due to missing data in longitudinal measurements, and to accommodate observations at irregularly spaced time points, we propose a smoothed composite likelihood approach for estimations. The forward intensity function approach intrinsically incorporates the future dynamics in the predictor variables that affect the stochastic occurrence of the future event. Thus the proposed framework is advantageous and parsimonious from requiring no separated modeling step for the stochastic mechanism of the predictor variables. Our theoretical analysis establishes the validity of the forward intensity modeling approach and the smoothed composite likelihood method. Extensive simulations and real-data analyses demonstrate the promising performance of this predictive approach.

EO100 Room Senate EXTREMES OF RANDOM GRAPHS AND GAUSSIAN FIELDS Chair: Juan Juan Cai

E1417: **Extremes of Gaussian random interfaces**
Presenter: **Alessandra Cipriani**, TU Delft, Netherlands

Random interfaces arise naturally as separating surfaces between two different thermodynamic phases or states of matter, for example oil and water.

When one views them as a field of random heights, it is natural to investigate the behavior of their rescaled maxima. We will review the current state of the art concerning the limiting behavior of extrema of Gaussian models that play a major role in statistical mechanics: the discrete Gaussian free field (DGFF), the membrane model (MM), and the $(\nabla+\Delta)$-model, which represents a mixture between DGFF and MM. We will present their similarities and differences, and show that the study of their extrema is connected with the theory of partial differential equations and numerical analysis.

E1444: **Asymptotic study of extremes in branching random walk**
Presenter: **Ayan Bhattacharya**, Wroclaw University of Science and Technology, Poland

Discrete time branching random walk on real line is considered where the displacements have regularly varying tail. Due to heavy-dependence structure and non-stationarity among the positions of the particles in the nth generation, it is challenging to study the growth of extreme positions. We shall discuss asymptotics for the bulk behavior of t extreme positions. As a consequence of heavy-dependence structure, a non-trivial strong cluster structure will appear in the limit.

E1337: **Extremes of stationary random fields on a lattice**
Presenter: **Chengxiu Ling**, Xián Jiaotong-Liverpool University, China

Extremal behavior of stationary Gaussian sequences/random fields is widely investigated since it models common cluster phenomena and brings a bridge between discrete and continuous extremes. We establish extensively limit theorems of stationary random fields under certain mixing and dependence conditions, which are further illustrated by typical examples of order statistics of Gaussian random fields and skew-Gaussian random fields. The positivity of the cluster index involved and its link with the expected cluster size are discussed.

EO178 Room CLO 204 NEW STATISTICAL APPROACHES FOR IMAGING AND DIGITAL DATA Chair: Taps Maiti

E0383: **Large-scale constrained joint modeling with applications to freemium mobile games**
Presenter: **Gourab Mukherjee**, University of Southern California, United States

A Constrained Extremely Zero Inflated Joint (CEZIJ) modeling framework is developed for simultaneously analyzing player activity, engagement and drop-outs in app-based mobile freemium games. The proposed framework addresses the complex interdependencies between a player's decision to use a freemium product, the extent of her direct and indirect engagement with the product, and her decision to permanently drop its usage. CEZIJ extends the existing class of joint models for longitudinal and survival data in several ways. It not only accommodates extremely zero-inflated responses in a joint model setting, but also incorporates domain-specific, convex structural constraints on the model parameters. Longitudinal data from app-based mobile games usually exhibit a large set of potential predictors and choosing the relevant set of predictors is highly desirable for various purposes, including improved predictability. To achieve this goal, CEZIJ conducts simultaneous, coordinated selection of fixed and random effects in high-dimensional penalized generalized linear mixed models. For analyzing such large-scale datasets, variable selection and estimation is conducted via a distributed computing based split-and-conquer approach that massively increases scalability and provides better predictive performance over competing predictive methods.

E0931: **Geometric and invariant aspects of complex functional data**
Presenter: **Karthik Bharath**, University of Nottingham, United Kingdom

High-resolution imaging has led to increasing availability of functional data with rich geometric information: range and/or domain might be manifolds with symmetries; or, set of functions themselves might have a nonlinear structure. We will consider some specific examples (e.g. closed 2D curves representing tumour shapes) of such data, discuss their common features, and present some of the challenges involved in developing statistical models and methods that are compatible with the underlying geometry and invariances.

E1466: **Tensor-on-tensor regression**
Presenter: **Eric Lock**, University of Minnesota, United States

In neuroimaging analysis and other fields, both predictors and outcomes can take the form of a multi-way array (i.e., a tensor). We propose a framework for the linear prediction of a multi-way array from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (L_2) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to metabolite resonance spectroscopy data.

EO440 Room SH349 RECENT ADVANCES FOR COMPLEX DATA ANALYSIS Chair: Juan Romo

E1828: **Stringing via manifold learning: Improving the functional representation of high-dimensional data**
Presenter: **Harold A Hernandez**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, M Carmen Aguilera-Morillo

Stringing via manifold learning is discussed. This method maps general high-dimensional data to functional data. It assumes that the sample vectors are realizations of a smooth stochastic process, observed with a scrambled order of its components. The key ingredient, a reordering step, is fundamental before recovering the true underlying process generating the data. The earlier version of stringing (based on multidimensional scaling) is improved by incorporating manifold learning. With this proposal it is possible to recover non-linear relationships between predictors, resulting in a better reordering of the data and a more reliable functional representation. Through simulation studies it is shown that the proposed method outperforms the original approach. The potential of using stringing and functional modeling in the high-dimensional scenario is also addressed. Real data applications on gene expression and single-nucleotide polymorphisms (SNPs) arrays are presented.

E1912: **A new selection criterion for statistical home range estimation**
Presenter: **Amparo Baillo**, Universidad Autonoma de Madrid, Spain
Co-authors: Jose E Chacon

The home range of a specific animal describes the geographic surface where this individual spends most of the time while carrying out its usual activities (eating, resting, reproduction, ...). Although a well-established definition of this concept is lacking, there are a variety of home range estimators. We address the open question of choosing the "best" home range from a collection of them based on the same sample of locations. We introduce the penalized overestimation ratio, a numerical index to rank the estimated home ranges. The key idea is to balance the excess area covered by the estimator (with respect to the original sample) and a shape descriptor measuring the over-adjustment of the home range to the data. To our knowledge, apart from computing the home range area, our ranking procedure is the first one which is both applicable to real data and to any type of home range estimator. Furthermore, the analysis and optimization of the selection index provides a way to select the tuning parameters of the home range estimators. For illustration purposes, using R, firstly we apply the new procedure to a set of real locations of a Mongolian wolf. Secondly, we carry out a Monte Carlo study to compare the true home range with the estimated one selected by our selection proposal.

E1688: **From halfspace M-depth to multiple-output expectile regression**
Presenter: **Davy Paindaveine**, Universite libre de Bruxelles, Belgium
Co-authors: Abdelaati Daouia

Despite the importance of expectiles in fields such as econometrics, risk management, and extreme value theory, expectile regression—or, more generally, M-quantile regression—unfortunately remains limited to single-output problems. To improve on this, we define hyperplane-valued multivariate M-quantiles that show strong advantages over their point-valued competitors. Our M-quantiles are directional in nature and provide centrality regions when all directions are considered. These regions define new statistical depths, the halfspace M-depths, that include the celebrated Tukey depth as a particular case. We study thoroughly the proposed M-quantiles, halfspace M-depths, and corresponding regions. M-depths not only provide a general framework to consider Tukey depth, expectile depth, L_r-depths, etc., but are also of interest on their own. However, since our original motivation was to consider multiple-output expectile regression, we pay more attention to the expectile case and show that expectile depth and multivariate expectiles enjoy distinctive properties that will be of primary interest to practitioners: expectile depth is maximized at the mean vector, is smoother than the Tukey depth, and exhibits surprising monotonicity properties that are key for computational purposes. Finally, our multivariate expectiles allow defining multiple-output expectile regression methods.

EC805 Room MAL 152 CONTRIBUTIONS IN NON- AND SEMI-PARAMETRIC STATISTICS Chair: Wenceslao Gonzalez-Manteiga

E0533: **Bootstrap bandwidth selection for matching estimators in nonparametric regression**
Presenter: **Ines Barbeito**, University of A Coruna, Spain
Co-authors: Stefan Sperlich, Ricardo Cao

The smoothed bootstrap method has been used in the context of prediction, in which the response variable of the target population remains unknown. Specifically, this bootstrap procedure is used for the purpose of bandwidth selection in regression estimation. The aim is to establish a new bootstrap bandwidth selector based on the exact expression of the bootstrap version of the mean average squared error of some approximation of the kernel regression estimator. This is very useful since Monte Carlo approximation is avoided for the implementation of the bootstrap selector. Furthermore, the distribution of the target population no longer needs to be estimated. The method is illustrated by applying it to a real data set which accounts for the gross annual income of men and women in Spain.

E1587: **Additive regression with metric-spaced-valued predictors and Hilbertian responses**
Presenter: **Jeong Min Jeon**, KU Leuven, Belgium
Co-authors: Byeong Park, Ingrid Van Keilegom

Nonparametric additive regression with general metric-space-valued predictors and general Hilbertian responses is considered. We estimate the component maps of the additive models by the smooth backfitting method. We present a general asymptotic theory and apply the theory to finite-dimensional Hilbertian predictors and manifold-valued predictors. Those predictors include not only standard Euclidean predictors but also compositional, circular, spherical and shape predictors. Our numerical study shows its wide applications.

E1608: **Missing mass estimation in feature sampling**
Presenter: **Federico Camerlenghi**, University of Milano-Bicocca and Collegio Carlo Alberto, Italy
Co-authors: Fadhel Ayed, Marco Battiston, Stefano Favaro

Feature models generalize species sampling models by allowing every observation to belong to more than one species, now called features. These models are extremely popular in machine learning and they found applications in diverse areas (e.g. in biosciences, biology and many others). Given a sample of size n, a relevant statistical problem related to these models is the estimation of the conditional expected number of hitherto unseen features that will be displayed in a future observation. Such a problem is usually referred to as the missing mass problem. This is motivated by numerous applied problems where the sampling procedure is expensive, in terms of time and/or financial resources allocated, and further samples can be only motivated by the possibility of recording new unobserved features. We introduce a simple, robust and theoretically sound nonparametric estimator of the missing mass, giving provable guarantees for its performance, and we derive corresponding confidence intervals via useful concentration inequalities. Our approach is illustrated through the analysis of various synthetic data and SNP data from the ENCODE sequencing genome project.

EG111 Room MAL 153 CONTRIBUTIONS IN CHANGE-POINTS Chair: Davide Giraudo

E1612: **The estimation of the traffic density and vehicle speed for detecting anomalies**
Presenter: **Kai Kasugai**, Chuo University, Japan
Co-authors: Toshinari Kamakura

It is an important issue to predict traffic density by using highway driving history data in order to ensure road safety on the highway. However, vehicles position and speed data often fail to represent the actual driving situation due to observation errors. It is often judged as abnormal even though an anomaly has not occurred. The purpose is to predict the state of the traffic density and vehicles speed at each point from the observed traffic density and vehicles speed at each point. With the Ensemble Kalman Filter, we could estimate precisely the traffic density and vehicles speed at each location following the observed values well. In case of calculating the Kalman gain, the stabilized results were obtained by applying a square root filter to the variance-covariance matrix. The proposed method illustrates easiness of detecting change points of traffic density and vehicles speed which results in findings of anomalies in the traffic highway system.

E1619: **Change-point regression robustified to smooth artifacts**
Presenter: **Florian Pein**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah, Paul Fearnhead

Piecewise constant regression is undeniable the most common form of change-point regression. While the assumption of a piecewise constant signal is reasonable in many applications, there are import examples, for instance genome sequencing to detect copy number variations or ion channel recordings (experiments to measure the conductance of a single ion channel over time), where such an assumption is at least questionable. Instead, such experiments can be better modelled by a piecewise constant function plus a smooth function. Existing methods are not using such a model explicitly for detecting change-points. Contrarily, they assume a piecewise smooth signal, either explicitly or a decomposition into piecewise constant plus smooth functions is stated but not used for detecting change-points, instead only the smooth function is reestimated in a second step. To use the decomposition explicitly, a modified fused lasso combined with smoothing techniques is proposed. Simulations show that this leads often to a better detection power. Moreover, the new methodology is very flexible. Kernel regression, smoothing splines and other methodologies can be used for smoothing. Secondly, extensions to multivariate and to filtered datasets are straightforward. Both extensions will be used to analyse genome sequencing data and ion channel recordings, respectively.

E1704: **Structural change detection via common principal component analysis**
Presenter: **Tatsuya Matsukawa**, Chuo University, Japan
Co-authors: Toshihiro Misumi, Yoshihiko Maesono, Sadanori Konishi

The problem of detecting a covariance structure change for multivariate data collected over time is considered. In order to detect the structure change point, we propose a new approach based on common principal component analysis (CPCA). The CPCA enables us to visually capture a pattern extraction and structural change as time passes on common principal component axes by simultaneously diagonalizing the variance-covariance matrices of all groups. To objectively identify the covariance structure change, we introduce a new procedure based on a deviance for the CPCA

model. Using the difference of the deviances between the full model and submodel under each common principal component assumption, we detect the change point of the covariance structure. We can also use the new method to reduce the dimension of the data. We investigate the effectiveness of our proposed method through the analysis of real data and a simulation study.

Saturday 14.12.2019 11:25 - 13:05 Parallel Session D – CFE-CMStatistics

EI012 Room Beveridge Hall STATISTICAL ANALYSIS OF NETWORKS Chair: Chenlei Leng

E0170: **'Statistics 101' for network data objects**
Presenter: **Eric Kolaczyk**, Boston University, United States

It is becoming increasingly common to see large collections of network data objects – that is, data sets in which a network is viewed as a fundamental unit of observation. As a result, there is a pressing need to develop network-based analogues of even many of the most basic techniques already standard for scalar and vector data. At the same time, principled extensions of familiar techniques to this context are nontrivial, given that networks are inherently non-Euclidean. We will present a number of results extending the notion of asymptotic inference for means to the contexts of various types of networks, i.e., both labeled and unlabeled, and either single- or multi-layer. These results rely on a combination of tools from geometry, probability theory, and statistical shape analysis. We will illustrate drawing from various applications in bioinformatics, computational neuroscience, and social network analysis under privacy.

E0172: **Generative link prediction for incomplete networks with node features**
Presenter: **Ji Zhu**, University of Michigan, United States

Link prediction is one of the fundamental problems in network analysis. Most existing methods require at least partial observation of connections for every node. In real-world networks, however, there often exist nodes that do not have any link information, and it is imperative to make link predictions for such nodes based on their node features. In this talk, we consider a general framework in which a network consists of three types of nodes: nodes having features only, nodes having link information only, and nodes having both. The goal is to predict links between nodes having features only and all other nodes. Under this setting, we have proposed a family of generative models for incomplete networks and node features, and we have developed a variational auto-encoder algorithm for model estimation and link prediction and investigated different encoder structures. We have also designed a cross-validation scheme under the problem setting. The proposed method has been evaluated on an online social network and two citation networks and achieved superior performance comparing with existing methods.

E0171: **Something old, something new, something borrowed, something not BLUE**
Presenter: **Karl Rohe**, University of Wisconsin-Madison, United States

A modern interpretation of a vintage technique in multivariate analysis will be given. This new perspective constitutes a fundamental advance in multivariate statistics (and spectral data analysis) that will blow your mind.

EO296 Room CLO B01 ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Naisyin Wang

E1747: **Predictive functional linear models with semiparametric single-index interactions**
Presenter: **Naisyin Wang**, University of Michigan, United States

When building a predictive model using both functional and multivariate predictors, it is often crucial to include the interaction between the two sets of predictors. To overcome the curse of dimensionality, we assume the interaction depends on a nonparametric, single-index structure of the multivariate predictor, and we reduce the dimensionality of the functional predictor using functional principal component analysis. We fit the model using an iterative procedure by minimizing a local quasi-likelihood using truncated FPCA series. By treating the number of FPCA scores as a tuning parameter and allowing it to diverge to infinity, we show that for a wide range of this truncation number and different bandwidths used by the nonparametric component in the single-index interaction, the parametric component of the model is root-n consistent and asymptotically normal. In addition, the overall prediction error is dominated by the estimation of the nonparametric function in the single-index interaction: an outcome that leads to a CV-based procedure to select the tuning parameters. We also show that the prediction error in the functional effect enjoys the minimax optimal rate. In a crop yield prediction application, we show that our single-index interaction model yields a lower prediction error than the conventional functional linear model and other competing nonlinear functional regression models.

E1105: **Functional structural equation modeling with high dimensional data in medical research**
Presenter: **Yuko Araki**, Shizuoka University, Japan

Structural equation modeling is a widely used multivariate statistical method to reveal structural relationships of several variables. In recent years in medical research, imaging data such as Magnetic Resonance imaging, have been more sophisticated and it its known that they give us important information. We propose a functional structural equation modeling to analyze association among several biomarkers including high dimensional image data. For effective and stable parameter estimation, the high dimensional data are expanded by using basis functions and sparse principal component analysis. We also discuss model selection methods based on information criterion. The proposed method is investigated through Monte Carlo simulations and applied to a medical study.

E0861: **Clustering and forecasting multiple functional time series**
Presenter: **Chen Tang**, The Australian National University, Australia
Co-authors: Han Lin Shang, Yanrong Yang

Modeling and forecasting age-specific mortality rates of multiple countries jointly could lead to improvements in long-term forecasting. Yet data that fed into joint models are often grouped according to nominal attributes, such as geographic regions, ethnic groups, and socio-economic status, which may still contain heterogeneity and thus deteriorate the forecast results. To address this, we propose a novel clustering technique to pursue homogeneity among multiple functional time series. Using functional panel data model with fixed effects, we are able to extract common features of functional time series. These common features could be decomposed into the functional time-trend and the mode of variations of functions. The proposed clustering method searches for homogeneous age-specific mortality rates of multiple countries by accounting for modes of variations and the temporal dependencies. Through simulation studies, we demonstrate that our proposed clustering technique outperforms other existing clustering methods, unless the objects are very similar. In a data analysis, we find that the clustering results of age-specific mortality rates can be explained by the combination of the aforementioned nominal attributes. We further show that our model produces better long-term forecasts than several benchmark methods for forecasting age-specific mortality rates.

E0864: **Multiple changepoints detection for a functional data sequence**
Presenter: **Yu-Ting Chen**, National Chengchi University, Taiwan
Co-authors: Jeng-Min Chiou

To detect multiple changes in a sequence of functional data, an algorithm called simultaneously dynamic segmentation (SDS) is proposed which performs dynamic segmentation simultaneously concerning different initial segments. SDS is free from the at-most-one-changepoint assumption and searches for the global minimum of the derived criterion. We also derive the null distribution of the objective function to determine the number of changepoints either forwardly or backwardly. We demonstrate the flexibility and validity of SDS through a simulation study and apply SDS to a traffic data set to find locations of changes in traffic conditions.

EO837 Room MAL B02 EARLY STOPPING RULES Chair: Alain Celisse

E1430: **Smoothed-residual stopping for statistical inverse problems via truncated SVD estimation**
Presenter: **Bernhard Stankewitz**, Humboldt University of Berlin, Germany

The purpose is to examine under what circumstances adaptivity for truncated SVD estimation can be achieved by an early stopping rule based on the smoothed residuals $\|A^{\alpha}(Y - A\hat{\mu}^{(m)})\|^2$. Lower and upper bounds for the risk are derived, which show that moderate smoothing of the residuals can be used to adapt over classes of signals with varying smoothness, while over-smoothing yields suboptimal convergence rates. The theoretical results are illustrated by Monte-Carlo simulations.

E1440: **Early stopping vs late stopping: Different flavors of SGD**
Presenter: **Nicole Muecke**, Institute for Stochastics and Applications, Germany

While stochastic gradient descent (SGD) is a workhorse in machine learning, the learning properties of many variants used in practise are hardly known. We consider non-parametric regression with (strongly) convex objectives and contribute to fill this gap focusing on the effect and interplay of multiple passes, mini-batching and averaging, and in particular tail averaging. An important aspect is choosing in a data-driven way the total number of iterations and the step-size, namely in terms of the localized empirical Rademacher Complexity. The results show how these different flavors of SGD can be combined to achieve optimal learning errors, providing also practical insights.

E1842: **Smoothed discrepancy principle as an early stopping rule in RKHS**
Presenter: **Yaroslav Averyanov**, Inria Lille-Nord Europe, France

The focus is on the estimation of a regression function that belongs to a reproducing kernel Hilbert space (RKHS). We describe spectral filter framework for our estimator that allows us to deal with several iterative algorithms: gradient descent, kernel ridge regression, etc. The main goal is to propose a new early stopping rule by introducing smoothing parameter for empirical risk of the estimator in order to improve the previous results on discrepancy principle. We explain, as well, how to generalise our strategy to different learning algorithms such as kNN and kernel regressions for choosing the hyperparameters. Theoretical justifications as well as simulations experiments for the proposed rule are provided.

E1955: **Improving on discrepancy-based early stopping with Tikhonov smoothing**
Presenter: **Alain Celisse**, Lille University, France
Co-authors: Martin Wahl

The purpose is to describe the early-stopping challenge and illustrate some deficiencies of classical discrepancy-based stopping rules. This motivates considering smoothing-based strategies such as the one inspired from Tikhonov regularization. For this rule, we prove several theoretical (non-)asymptotic guarantees, and also illustrate its promising practical behavior on simulation experiments carried out by means of spectral filter algorithms.

EO665 Room MAL B04 ADVANCES IN NETWORK AND MATRIX DATA ANALYSIS Chair: David Choi

E1360: **An omnibus embedding for multiple random graphs, and applications to multiscale joint network inference**
Presenter: **Avanti Athreya**, Johns Hopkins University, United States

Principled, scalable methods for statistical inference across multiple graphs is of vital importance in a host of application domains. We describe an omnibus embedding in which multiple graphs on the same vertex set are jointly embedded into a single space with a distinct representation for each graph. We prove a central limit theorem for this omnibus embedding, and we show that this simultaneous embedding into a single common space allows for the comparison of graphs without further orthogonal alignments. Moreover, the existence of multiple embedded points for each vertex renders possible the resolution of important multiscale graph inference goals, such as the identification of specific subgraphs or vertices as drivers of similarity or difference across large networks. We demonstrate the utility of the omnibus embedding in two analyses of connectomic graphs generated from MRI scans of the brain in human subjects. We show how the omnibus embedding can be used to detect statistically significant differences, at multiple scales, across these networks, with an identification of specific brain regions that are associated with population-level differences and which may be loci of markers of pathology.

E1403: **Asymptotically efficient estimators for stochastic blockmodels**
Presenter: **Minh Tang**, North Carolina State University, United States
Co-authors: Joshua Cape, Carey Priebe

Asymptotic normality results are established for estimation of the block probability matrix B in stochastic blockmodel graphs using spectral embedding when the average degrees grow at the rate of $\omega(n^{1/2})$ in n, the number of vertices. As a corollary, we show that when B is of full-rank, estimates of B obtained from spectral embedding are asymptotically efficient. When B is singular the estimates obtained from spectral embedding can have smaller mean square error than those obtained from maximizing the log-likelihood under no rank assumption, and furthermore, can be almost as efficient as the true MLE that assume known $rk(B)$. The results indicate, in the context of stochastic blockmodel graphs, that spectral embedding is not just computationally tractable, but that the resulting estimates are also admissible, even when compared to the purportedly optimal but computationally intractable maximum likelihood estimation under no rank assumption.

E1516: **Adaptive estimation of multivariate piecewise constant functions**
Presenter: **Sabyasachi Chatterjee**, University of Illinois at Urbana Champaign, United States

The estimation of multivariate piecewise constant functions is considered. A natural estimator is the Dyadic Cart estimator. We show that an extension of Dyadic Cart attains the best possible risk (up to log factors) adaptively for all piecewise constant functions in dimension 2. In higher dimensions, such a property continues to hold for a special subclass of all piecewise constant functions, but not all. Along the way, we show some new adaptive results for estimation of functions of bounded variation using dyadic cart and its extensions.

E1772: **Latent space representations of hypergraphs**
Presenter: **Simon Lunagomez**, Lancaster University, United Kingdom
Co-authors: Christopher Nemeth, Edoardo Airoldi, Kathryn Turnbull

The increasing prevalence of relational data describing interactions among a target population has motivated a wide literature on statistical network analysis. In many applications, interactions may involve more than two members of the population and this data is more appropriately represented by a hypergraph. We present a model for hypergraph data which extends a previous latent space distance model and, by drawing a connection to constructs from computational topology, we develop a model whose likelihood is inexpensive to compute. We obtain posterior samples via an MCMC scheme and we rely on Bookstein coordinates to remove the identifiability issues associated with the latent representation. We demonstrate that the latent space construction imposes desirable properties on the hypergraphs generated in our framework and provides a convenient visualisation of the data. Furthermore, through simulation, we investigate the flexibility of our model and consider estimating predictive distributions. Finally, we explore the application of our model to a real world co-occurrence dataset.

EO731 Room MAL B18 BIG BIAS IN BIG DATA: CAN WE CORRECT? Chair: Patrice Bertail

E1736: **Economic biases and algorithms**
Presenter: **Patrick Waelbroeck**, Telecom Paris, France
Biases in algorithms that result from cognitive biases and economic biases are discussed. A review of the main sources of biases is presented and the mechanism through biases related to economic strategies arise are analysed. Solution based on semi-structural modelling are compared with both reduced form and full structural estimations in econometrics.

E0631: **Nonparametric estimation for big-but-biased data**
Presenter: **Laura Borrajo**, UDC, Spain
Co-authors: Ricardo Cao
Nonparametric estimation for large-sized samples subject to sampling bias is studied. The general parameter considered is the mean of a transformation of the random variable of interest. When ignoring the biasing weight function, a small-sized simple random sample of the real population is assumed to be additionally observed. A new nonparametric estimator that incorporates kernel density estimation is proposed. Asymptotic properties for this estimator are obtained under suitable limit conditions on the two sample sizes and standard and non-standard asymptotic conditions on the two bandwidths. Explicit formulas are shown for the particular case of mean estimation. Simulation results show that the new mean estimator outperforms two classical ones. The influence of two smoothing parameters on the performance of the final estimator is also studied, exhibiting a striking behavior. The new method is applied to a real data set concerning airline on-time performance of US flights.

E0437: **Weighted empirical risk minimization: Transfer learning based on importance sampling**
Presenter: **Charles Tillier**, Telecom ParisTech, France
Co-authors: Stephan Clemencon, Robin Vogel, Mastane Achab
Statistical learning problems are considered when the distribution P' of the training observations $Z'_1, \ldots, Z'_n$ differs from the distribution P involved in the risk one seeks to minimize (referred to as the test distribution) but is still defined on the same measurable space as P and dominates it. In the unrealistic case where the likelihood ratio $\Phi(z) = dP/dP'(z)$ is known, one may extend the Empirical Risk Minimization (ERM) approach to this specific transfer learning setup using the same idea as that behind Importance Sampling, by minimizing a weighted version of the empirical risk functional computed from the 'biased' training data Z'_i with weights $\Phi(Z'_i)$. Although the importance function $\Phi(z)$ is generally unknown in practice, in various situations frequently encountered in practice, it takes a simple form and can be directly estimated from the Z'_i's and some auxiliary information on the statistical population P. Besides, we will see that the generalization capacity of the approach aforementioned is preserved when plugging the resulting estimates of the $\Phi(Z'_i)$'s into the weighted empirical risk.

E0465: **Empirical risk minimization under random censorship**
Presenter: **Guillaume Ausset**, Telecom ParisTech, France
Co-authors: Stephan Clemencon, Francois Portier
A classic supervised learning problem is considered, where a continuous non-negative random label Y (i.e. a random duration) is to be predicted based upon observing a random vector X valued in $\mathbb{R}^d$ with $d \geq 1$ by means of a regression rule with minimum least square error. In various applications, ranging from industrial quality control to public health through credit risk analysis for instance, training observations can be right censored, meaning that, rather than on independent copies of (X,Y), statistical learning relies on a collection of $n \geq 1$ independent realizations of the triplet $(X,\ \min\{Y,\ C\},\ \delta)$, where C is a nonnegative r.v. with unknown distribution, modeling censorship and $\delta = \mathbb{I}\{Y \leq C\}$ indicates whether the duration is right censored or not. As ignoring censorship in the risk computation may clearly lead to a severe underestimation of the target duration and jeopardize prediction, we propose to consider a plug-in estimate of the true risk based on a Kaplan-Meier estimator of the conditional survival function of the censorship C given X, referred to as Kaplan-Meier risk, in order to perform empirical risk minimization. It is established, under mild conditions, that the learning rate of minimizers of this biased/weighted empirical risk functional is of order $O_{\mathbb{P}}(\sqrt{\log(n)/n})$ when ignoring model bias issues inherent to plug-in estimation, as can be attained in absence of censorship.

EO302 Room MAL B20 INFERENCE ON CAUSAL PARAMETERS USING MACHINE LEARNING Chair: Xavier de Luna

E1768: **The costs and benefits of valid inference on causal parameters in the presence of high dimensional nuisance parameters**
Presenter: **Niloofar Moosavi**, Umea university, Sweden
Co-authors: Xavier de Luna, Jenny Haggstrom
The purpose is to study estimators yielding valid inference on a low dimensional causal parameter in the presence of high dimensional nuisance parameters. Naive estimation strategies based on regularisation or a preliminary model selection stage have finite sample distributions which are badly approximated by their asymptotic distributions. To solve this problem estimators which converge uniformly in distribution over a class of DGPs allowing for the number of parameters to increase with the number of observations, have recently been proposed in the literature. Uniform asymptotic results guarantee valid inference. However, this is often obtained at the cost of variance inflation, which can be severe in some settings as our simulation results show. In particular, these procedures may yield unnecessarily wide confidence intervals. We present an estimator which converges uniformly and thereby yields valid inference, but whose implied cost in terms of variability inflation is much lower than others. Approximate sparsity conditions are necessary as in the earlier literature.

E1008: **Robust double machine learning inference for conditional exposure effects**
Presenter: **Stijn Vansteelandt**, Ghent University and London School of Hygiene and Tropical Medicine, Belgium
Co-authors: Oliver Dukes
The evaluation of exposure effects from observational studies typically requires adjustment for high-dimensional confounding. This makes standard parametric inferences not entirely satisfactory as model misspecification is likely, and even relatively minor misspecifications over the observed data range may induce large bias in the exposure effect estimate. Over the past 2 decades, there has therefore been growing interest in the use of machine learning methods to assist this task. Naive use of machine learning is itself problematic as the resulting exposure effect estimate is prone to a so-called plug-in bias, and the bootstrap is not guaranteed to deliver valid confidence intervals. Pioneering work on Targeted Maximum Likelihood Estimation, and more recently on Double Machine Learning, has shown how this can be overcome by relying on double robust estimators. In particular, valid inference can be obtained by cleverly relying on machine learning predictions of both exposure and outcome, provided that both algorithms converge sufficiently fast to the truth. We will adapt so-called bias-reduced double-robust estimators to ensure valid inference even when one of the machine learning algorithms does not converge (fast) to the truth, thereby yielding results that deliver better approximations in moderate sample sizes.

E0371: **Collaborative inference for causal effect estimation and general missing data problems**
Presenter: **David Benkeser**, Emory University, United States
Doubly robust estimators are a popular means of estimating causal effects. Such estimators combine an estimate of the conditional mean of the study outcome given treatment and confounders (the so-called outcome regression) with an estimate of the conditional probability of treatment given confounders (the propensity score) to generate an estimate of the effect of interest. They have several desirable features: they are consistent so long as at least one of these two regressions is consistently estimated; they are also often efficient and achieve the lower bound on the variance of regular,

asymptotically linear estimators. Moreover, they facilitate the use of modern machine learning methods in the estimation of the outcome regression and propensity score. However, in problems where causal parameters are weakly identified, doubly robust estimators may behave erratically. We propose a new framework for inference in these challenging settings. We introduce the idea of collaborative asymptotically linear estimators. These estimators use doubly robust frameworks; however, rather than using an estimate of the propensity score, they opt for an alternative quantity with reduced dimension relative to the true propensity score. We will discuss these issues in the context of estimating the causal effect of a binary treatment on an outcome, and discuss extensions to a general setting where the observed data are the result of a conditionally random coarsening of a full data structure.

E1396: **Sensitivity analysis via the proportion of unmeasured confounding**
Presenter: **Matteo Bonvini**, Carnegie Mellon University, United States
Co-authors: Edward Kennedy

In observational studies, identification of average treatment effects is generally achieved by assuming no unmeasured confounding, possibly after conditioning on enough covariates. Because this assumption is both strong and untestable, a sensitivity analysis should be performed. Widely used approaches include modeling the bias directly or varying the propensity scores to probe the effects of a potential unmeasured confounder. We take a novel approach whereby the sensitivity parameter is the proportion of unmeasured confounding. We consider several scenarios imposing different assumptions on the probability of a unit being unconfounded. In each case, we derive sharp bounds on the average treatment effect as a function of the sensitivity parameter and propose nonparametric estimators. We introduce a one-number summary of a study's robustness to the number of confounded units. Finally, we explore finite-sample properties via simulation, and apply the methods to an observational database used to estimate the effects of Right Heart Catheterization (RHC) on the care of critically ill patients.

EO747 Room MAL B35 MARKED RECURRENT EVENT PROCESSES WITH INCOMPLETE OBSERVATIONS Chair: Sy Han Chiou

E1663: **Covariate balancing with recurrent marker data**
Presenter: **Kwun Chuen Gary Chan**, University of Washington, United States
Co-authors: Raymond Wong

A nonparametric estimation method is discussed for the counterfactual cumulative rate functions, for a binary exposure. The weights are constructed through balancing covariate distributions between each exposure category and the combined data. Covariate balance is often advocated for objective causal inference since it mimics randomization in observational data. Unlike methods that balance specific moments of covariates, the proposal attains uniform approximate balance for covariate functions in a reproducing-kernel Hilbert space. We will discuss connection between the corresponding infinite-dimensional optimization problem to a finite-dimensional eigenvalue optimization problem. An advantage of such weighting is that it can be applied to multiple functional components in the estimation of cumulative rate functions.

E1789: **Analyzing wearable device data using marked point processes**
Presenter: **Yuchen Yang**, Johns Hopkins University, United States
Co-authors: Mei-Cheng Wang

Two sets of measures are introduced as exploratory tools to study physical activity patterns: active-to-sedentary/sedentary-to-active rate function (ASRF/SARF), and active/sedentary rate function (ARF/SRF). These two sets of measures are complementary to each other and can be effectively used together to understand physical activity patterns. The specific features are illustrated by an analysis of wearable device data from National Health and Nutrition Examination Survey (NHANES). A two-level semiparametric regression model for ARF and the associated activity magnitude is developed under a unified framework using the marked point process formulation. The inactive and active states measured by accelerometers are treated as 0-1 point process, and the activity magnitude measured at each active state is defined as a marked variable. The commonly encountered missing data problem due to device non-wear is referred to as "window censoring", which is handled by a proper estimation approach that adopts techniques from recurrent event data. Large sample properties of the estimator and comparison between two regression models as measurement frequency increases are studied. Simulation and NHANES data analysis results are presented. The statistical inference and analysis results suggest that ASRF/SARF and ARF/SRF provide useful analytical tools to practitioners for future research on wearable device data.

E1994: **Evaluating the effect of interventions on recurrent event outcomes with multiple competing risks**
Presenter: **Elizabeth Colantuoni**, Johns Hopkins University, United States

Among older adults who are critically ill and receiving care in an intensive care unit (ICU), delirium is very common (>70%) and strongly associated with long-term cognitive impairment that is similar to Alzheimers Disease and Related Dementias (ADRD). Hence, there is a growing number of clinical trials evaluating interventions to prevent or treat delirium, with the goal of reducing associated long-term cognitive impairment/ADRD in critical illness survivors. Evaluating the impact of interventions on delirium is complicated due to the time-varying nature of delirium, the fact that delirium can frequently re-occur, and patients may experience death or be discharged from the ICU prior to completing the pre-planned duration of patient follow-up (e.g. 14 days). In addition, it is often hypothesized that interventions that may reduce delirium onset or duration would also reduce mortality and duration of ICU stay. Current approaches to evaluate the delirium endpoint rely on aggregating the data into a simple binary summary (ever vs. never delirium) or a composite endpoint (delirium free days to 14 days). These approaches will be contrasted to a novel extension of an existing method, the joint recurrent and terminating event model, that allows for the multiple terminating events of death and ICU discharge.

E2017: **The use of joint modelling to analyze recurrent events: Appearance of new lesions in cancer.**
Presenter: **Virginie Rondeau**, Inserm, France
Co-authors: denis rustand, agnieszka Krol

The Response Evaluation Criteria in Solid Tumors are used as standard guidelines for the clinical evaluation of cancer treatments. The assessment is based on the anatomical tumor burden: change in size of target lesions and evolution of nontarget lesions (NTL). Despite unquestionable advantages of this standard tool, Response Evaluation Criteria in Solid Tumors are subject to some limitations such as categorization of continuous tumor size or negligence of its longitudinal trajectory. In particular, it is of interest to capture its nonlinear shape and model it simultaneously with recurrent progressions of NTL and overall survival. We propose different multivariate nonlinear joint frailty models for recurrent events, and a terminal event with a longitudinal data. In the model, the tumor size trajectory accounts for the natural growth and treatment induced decline. The semicontinuous nature of the longitudinal biomarker is specified by a two-part model, which splits its distribution into a binary outcome represented by the positive versus zero values and a continuous outcome (second part) with the positive values only. On real data on colorectal cancers, we determine on which component, tumor size, NTL, or death the treatment acts mostly and perform dynamic predictions of death.

EO050 Room MAL B36 ROBUST MACHINE LEARNING Chair: Adrien Saumard

E0296: **Polynomial-time estimation of the mean**
Presenter: **Olivier Collier**, Universita Paris-Nanterre, France

Independent Gaussian observations in high dimension are considered which share the same mean for the most part, while the other means can be arbitrarily large. We study the problem of robustly estimating the common mean in the minimax sense. But more precisely, we aim at finding feasible procedures, i.e. computable in polynomial time. First, we show the relation between robustly estimating the mean and estimating some

linear functionals of the outliers. Then, we define a group-LASSO-like procedure for estimating the mean, which has better performance as previously existing methods. However, computational tractability comes with a loss of minimax-rate-optimality in some regimes.

E1154: **Robust clustering algorithm based on Median-of-Means statistics**
Presenter: **Edouard Genetay**, CREST-ENSAI, France
Co-authors: Adrien Saumard, Camille Saumard

Classical clustering methods, such as K-means, suffer from a lack of robustness with respect to outliers. We propose a robust version of K-means, using median-of-means statistics, a strategy that has been recently put to emphasis for efficient robust machine learning. The algorithm is iterative, in a Lloyd-type fashion. We propose an efficient initalization and empirically show rapid convergence along the iteration steps. The algorithm clearly outperforms K-means on corrupted or heavy-tailed data and is competitive with other robust approaches, such as K-median for instance. As an additional outcome, our algorithm provides a detection of outliers.

E1413: **Langevin sampling for median of means based estimation**
Presenter: **Wenjuan Sun Stephane Chretien**, National Physical Laboratory - University of Lyon 2, United Kingdom
Co-authors: Stephane Chretien

Median of Means (MoM) estimators have attracted a lot of interest lately due to their ability to circumvent being corrupted by outliers and heavy tailed training data. However, designing fast algorithms for computing MoM estimators is still an interesting challenge both in practice and theory. We will show how the Langevin sampler can be put to work for MoM estimation. We show that the Langevin sampler reaches neighborhoods of the set of stationary points of the expected risk in polynomial time. We will also present experimental results for XCT reconstruction using deep learning based "learned gradient" methods.

E1891: **A noise-robust fast sparse Bayesian learning model**
Presenter: **Ingvild Margrethe Helgoy**, University of Bergen, Norway
Co-authors: Yushu Li

The hierarchical model structure from the Bayesian Lasso is utilized in the Sparse Bayesian Learning process to develop a new type of probabilistic supervised learning approach. This approach has several performance advantages, such as being fast, sparse and especially robust to the variance in random noise. The hierarchical model structure in this Bayesian framework is designed in such a way that the priors do not only penalize the unnecessary complexity of the model but also depend on the variance of the random noise in the data. The hyperparameters in the model are estimated by the Fast Marginal Likelihood Maximization algorithm and can achieve low computational cost and faster learning process. We compare our methodology with two other popular Sparse Bayesian Learning models: The Relevance Vector Machine and a sparse Bayesian model that has been used for signal reconstruction in compressive sensing. We show that our method will generally provide more sparse solutions and be more flexible and stable when data is polluted by high variance noise.

EO264 Room G3 MEASURES OF SYSTEMIC RISK IN ACTUARIAL SCIENCE AND FINANCE Chair: Zinoviy Landsman

E0212: **Risk management application of intrinsic discrepancy loss functions**
Presenter: **Udi Makov**, University of Haifa, Israel
Co-authors: Zinoviy Landsman, Limor Langbord

One of the main problems in risk management is the evaluation of risk measures, which are typically explored for their mathematical properties, under the assumption that all the underlying parameters of the loss distribution are known. In practice, little attention is given to the impact the choice of parameter estimates has on the accuracy of such measures. We propose to estimate these parameters by minimizing the expectation of the intrinsic discrepancy loss function (IDLF), which is the inherent loss function arising only from the underlying distribution, without any external subjective considerations. Firstly, we discuss the intrinsic estimation of the mean of the Tweedie family, which is a subclass of the reproductive exponential dispersion family, a reach family with wide applications in risk management, and provide the IDLF of the family. Secondly, we provide a numerical study of the estimates of the Value at Risk (VaR) and the Tail Conditional Expectation (TCE) of the gamma loss distribution using the IDLF and its approximation.

E0651: **A novel multi-elliptical family of distributions: Definitions, properties and risk capital decomposition**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Tomer Shushi

The multivariate elliptical family of distributions is well studied and commonly used in actuarial science and finance. However, it has an essential shortcoming: all its univariate marginal distributions are the same, up to location and scale transformations. This happens because these marginal distributions have the same density generator. For example, all marginals of the multivariate Student-t distribution, an important member of the elliptical class, have the same number of degree of freedoms. We introduce a generalization of the multivariate elliptical family of distributions that considers marginals with different density generators. This becomes important when dealing with insurance and financial data. We further provide the main characteristics of the multi-elliptical family of distributions: characteristic and density functions, expectations and covariance matrices. Furthermore, we derive important risk measures for the introduced distributions, such as the value at risk (VaR) and tail conditional expectation (TCE). We also provide the TCE-based capital allocation of aggregate risks.

E0666: **A novel approach to measure systemic risks using multivariate tail moments: From theory to practice**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel
Co-authors: Zinoviy Landsman, Udi Makov

Systemic risks have been proved to be extremely harmful to the Financial system with a potential for a catastrophic failure occurring when risks are mutually dependent. In practice, risk managers that focus on the possibility for a crisis are confronted with not only one risk but rather a system of risks (such as several business lines). So the world of risks is, in fact, multivariate and in this context dealing with univariate risk measures is inadequate. We will present a novel approach to building systemic risk measures based on multivariate tail moments. While having intuitive reasoning and clear theoretical foundations, they can also be directly applied for various problems of assessing the risk from a system of mutually dependent risks. In particular, we will present the multivariate tail conditional expectation and the multivariate tail covariance matrix, as natural extensions of the expected shortfall and tail variance measures, respectively. Then, several aspects will be examined, showing the capability of such an approach in practice.

E1023: **Downside risk optimization with random targets and portfolio amplitude**
Presenter: **Jing Yao**, Heriot-Watt University, United Kingdom

Using downside risk optimization subject to a random target in portfolio selection is rationalized. In context of normality, we derive analytical solutions to the downside risk optimization with respect to random targets and investigate how the random target affects the optimal solutions. In doing so, we propose using portfolio amplitude, as a new measure in literature, to characterize the investment strategy. Particularly, we demonstrate the mechanism by which the random target inputs its impact into the system and alters the optimal portfolio selection. Our results underpin why investors prefer holding some specific assets in following random targets and provide explanations for some special investment strategies, such as constructing a stock portfolio following a bond index. Numerical examples are presented to clarify our theoretical results.

EO166 Room G5 CSDA JOURNAL Chair: Ana Colubi

E1397: **Testing parameters of models with conditional moment restrictions using empirical likelihood**
Presenter: **Yves Berger**, University of Southampton, United Kingdom
An empirical likelihood test is proposed for models defined by conditional moment restrictions, such as models with non-linear endogenous covariates, with or without heteroscedastic errors or non-separable transformation models. The number of empirical likelihood constraints is given by the size of the parameter, unlike alternative approaches. We show that the empirical likelihood ratio test is asymptotically pivotal, without explicit studentisation. For moderate sample sizes, we show that this property may not hold with alternative empirical likelihood approaches. The proposed test clearly outperforms alternative empirical likelihood tests. It also offers a major advantages over two-stage least-squares, because the relationship between the endogenous and instrumental variables does not need to be known.

E1435: **Modelling the extremes of flu cases**
Presenter: **Eva Cantoni**, University of Geneva, Switzerland
Co-authors: Setareh Ranjbar, Valerie Chavez-Demoulin, Giampiero Marra, Rosalba Radice
From a health care managerial point of view, the number of tested patients showing flu-like symptoms and the number of positive (or negative) cases of the flu are important indicators of the epidemic of flu and of congestion in a hospital, respectively. We model the extremes of positive cases of flu, of suspicious cases to be tested as well as the ratio of the two. We analyse three years (2016 - 2019) of daily data from the University hospital (CHUV) in Lausanne (Switzerland) with the peak over threshold method. We fit a Discrete Generalized Pareto Distribution (DGPD) to the number of tested patients and number of positive cases, and a Generalized Pareto Distribution (GPD) to the ratio of positive to tested in the framework of Generalized Additive Model for Location, Shape and Scale, which allows full flexibility on the functional form between the covariates and the parameters of the distributions. Since the data show the potential presence of outlying observations, possibly related to the simultaneous outbreak of other diseases sharing the same symptoms with the flu, we develop robust methods to fit the DGPD and GPD. The robust method is crucial for providing a reliable prediction of the flu episodes. The robust data analysis show that the seasonality in the flu data is better explained by meteorological data rather than calendar dates. The reliability of the proposal is supported by an extensive simulation study.

E0831: **On the BHEP test for functional data**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain
Co-authors: Norbert Henze
Many statistical procedures for finite dimensional data assume the data To be normally distributed. Thus, a number of normality tests have been proposed. Some of them are based on comparing a nonparametric estimator of a function characterizing a probability law with a parametric estimator of that function, obtained under the null hypothesis. The BHEP test belongs to this class of tests. It is based on comparing the empirical characteristic function with the characteristic function of the normal law. The normality assumption is important not only in the classical context, but also in other settings such as functional data analysis. Since the probability measure of a random element taking values in a separable Banach space is characterized by its characteristic function, the objective is to extend the BHEP to the functional data context.

E1676: **Robust estimation of the Pickands dependence function under random right censoring**
Presenter: **Yuri Goegebeur**, University of Southern Denmark, Denmark
Co-authors: Armelle Guillou , Jing Qin
Robust nonparametric estimation of the Pickands dependence function under random right censoring is considered. The estimator is obtained by applying the minimum density power divergence criterion to properly transformed bivariate observations. The asymptotic properties are investigated by making use of results for Kaplan-Meier integrals. We investigate the finite sample properties of the proposed estimator with a simulation experiment and illustrate its practical applicability on a dataset of insurance indemnity losses.

EO258 Room MAL G13 SIGNAL PROCESSING Chair: Rosa Maria Fernandez-Alcala

E0750: **Change detection with quaternion random signals**
Presenter: **Antonia Oya**, Universidad de Jaen, Spain
The problem of change detection in continuous-time random signals using statistical tests has been applied for a number of situations, such as medical condition monitoring, climate change detection, speech and image analysis, stock market, traffic data analysis and so on. A typical statistical formulation of this detection problem can be regarded as discrimination between two probability distributions with different models, i.e., the probability distributions of data before and after a candidate change point. In these approaches, the logarithm of the likelihood ratio between two consecutive intervals in continuous-time signals is monitored for detecting change points. A test statistic for change detection with quaternion random signals based on Reproducing Kernel Hilbert Space (RKHS) formulation is given. Specifically, a change detection problem where the pre-change observation signal is purely noise and the post-change observation signal is a noise corrupted signal is considered. First, we reduce the continuous-time change detection problem to a discrete setting by considering the random coefficients obtained from the RKHS representation of the observation quaternion random signal. Second, we compute the log-likelihood ratio to obtain a feasible expression of the detector that involves RKHS inner product computations.

E1137: **Signal processing for statistical arbitrage**
Presenter: **Samuel Weller**, Royal Holloway University of London, United Kingdom
Co-authors: Clive Cheong Took
Statistical arbitrage (sometimes referred to as pairs trading) is a market neutral (open both a long and short position) trading strategy which creates profit from pricing inefficiencies between a pair of securities. Typically, this is done by creating mathematical models to determine the best course of action between two securities. Signal Processing and Machine Learning solutions which will be designed to analyse the relationship between a pair of given securities, and thereafter determine whether the pair of assets is worth trading or not.

E1148: **Estimation recursive algorithms of hypercomplex signals from delayed observations**
Presenter: **Jose Domingo Jimenez-Lopez**, University of Jaen, Spain
Co-authors: Jesus Navarro-Moreno
In many practical situations, as in communication mechanisms, robots, signal detectors, electronic devices, among others, intermittent failures in the signal transmission occur, and this fact causes errors in the observations. The state-space model staying within these situations includes Bernoulli random variables in the observation equation, which allows us to approach uncertain, delayed or lost observations. We study the one or two steps delayed observations case, that is, at each instant of time, every quaternion component of the available observation can coincide with the corresponding component of the real observation, or be delayed one or two sampling periods. Moreover, recently many real situations in fields as signal and image processing, quantum and orbital mechanics, graphics computer, attitude control or robotics, among others, are modelled by quaternion signals, because they need orientation and rotation characteristics to be modelled. We approach the linear processing of quaternion signals from delayed observations, proposing recursive algorithms for different estimation problems and considering the complete quaternion, that is, the real component with its three involutions.

E1086: **A SWL fixed-lag smoothing algorithm with intermittent observations in the presence of correlated noises**
Presenter: **Rosa Maria Fernandez-Alcala**, University of Jaen, Spain
Co-authors: Rahul Yadav

A new fixed-lag smoothing algorithm in the quaternion domain is devised in order to handle intermittent observations in the presence of correlated noises at the same time. The problem is formulated in the state-space framework where a sequence of independent Bernoulli random variables is introduced to model random interruptions. Assuming the quaternion signal is C-proper, the proposed methodology is based on a semi-widely linear processing which reduces the dimension of the problem to a half. In this approach, the initial state space-model involving autocorrelated and crosscorrelated noises is reformulated to make the noise sequence does not correlate with each. Subsequently, a similar reasoning that of the quaternion Kalman filtering equations is followed. The benefits of the proposed solution over the conventional fixed-lag smoothing algorithm are illustrated by means of a simulation example.

EO106 Room MAL G14 RECENT ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Marco Ferreira

E0658: **Fast and scalable posterior simulation for Gaussian hierarchical models with ICAR spatial random effects**
Presenter: **Marco Ferreira**, Virginia Tech, United States

A novel algorithm is developed for the simulation from the posterior distribution of Gaussian hierarchical models with intrinsic conditional autoregressive (ICAR) spatial random effects. ICAR specifications assume a neighborhood structure that implies a sparse precision matrix for the spatial random effects. The algorithm is based on the spectral decomposition of the spatial random effects precision matrix. The algorithm is a Markov chain Monte Carlo algorithm that scales linearly with the size of the sample size. We perform a simulation study that shows improved performance of our algorithm when compared to competing existing posterior simulators. Finally, we illustrate the application of our novel algorithm with a spatial regression study of median household income in the contiguous United States in 2017 per county.

E0726: **Dynamical mixture modelling of spatial processes**
Presenter: **Thais Fonseca**, Universidade Federal do Rio de Janeiro, Brazil
Co-authors: Alexandra Schmidt, Viviana Lobo

Spatio-temporal processes in environmental applications are usually assumed to follow a Gaussian model, possibly after some transformation. Flexibility to the usual Gaussian assumption is added by modelling the process as a scale mixture between a Gaussian and log-Gaussian process. The scale is represented by a process and it is allowed to depend on covariates. The resultant kurtosis varies with location, allowing the time series at each location to have different distributions with different tail behaviour. Regarding the temporal dependence, a dynamical model based on state equations is assumed for the scale process and a computationally efficient estimation algorithm based on sequential Monte Carlo is proposed. An application to maximum temperature data illustrates the effects of altitude in the variability of the process and how this dependence may change over time.

E0793: **Variable selection in the presence of factors: A model selection perspective**
Presenter: **Rui Paulo**, ISEG/CEMAPRE, Universidade de Lisboa, Portugal
Co-authors: Gonzalo Garcia-Donato

The variable selection problem where the set of potential predictors contains both factors and numerical variables is considered. There are two possible approaches to variable selection: the estimation-based and the model-selection-based. In the former, the model containing all the potential predictors is estimated and a criterion for excluding variables is devised based on the estimate of the associated parameters. In the latter, all 2^p models are considered, where p stands for the number of potential predictors, and variable selection is based on the posterior distribution on the model space. Inducing sparsity is a major challenge in the estimation-based approach, while the model-selection-based techniques are subject to the issue of multiplicity. We approach the variable selection problem in the presence of factors via the model selection perspective. Formally, this is a particular case of the standard variable selection setting where factors are coded using dummy variables. Nevertheless, we show several inputs like the assignment of prior probabilities over the model space or the parameterization adopted for factors may have a large (and difficult to anticipate) impact on the results. We provide a solution for these issues that extends the proposals in the standard variable selection problem and does not depend on how the factors are coded using dummy variables. Additionally, our method exhibits a very competitive frequentist behavior.

E1178: **Decoupling shrinkage and selection in Gaussian linear factor analysis**
Presenter: **Hedibert Lopes**, INSPER, Brazil
Co-authors: Henrique Bolfarine, Carlos Carvalho, Jared Murray

Sparsity-inducing priors has been a relevant option in variable selection in a variety of statistical models. Despite the interpretability and ease of application, there are still divergences in determining whether a parameter a posteriori is really null. In this context, the decoupling shrinkage and selection (DSS) approach appears as an alternative that preserves the a posteriori information while providing an optimal selection in the set of variables. We extend the DSS methodology for the Gaussian linear factor analysis model in order to obtain a sparse loadings matrix, reducing to zero the parameters that are not relevant to the model. To perform such selection, we introduce a penalized loss function, a post inference procedure that relies on a penalized predictive version of the expectation-maximization (EM) algorithm, and a graphical summary. The findings are illustrated with simulations and two applications, the first in psychometrics and the latter in denoising of handwritten data. The standard normal and point mass priors were used, resulting in significantly different levels of sparsity in the recovered loading matrix.

EO572 Room MAL G15 ADVANCES IN BAYESIAN MODELING AND MODEL SELECTION Chair: Raffaele Argiento

E0240: **Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data**
Presenter: **Alberto Cassese**, Maastricht University, Netherlands
Co-authors: Qiwei Li, Michele Guindani, Marina Vannucci

A Bayesian hierarchical mixture regression model is developed for studying the association between a multivariate response, measured as counts on a set of features, and a set of covariates. We have available RNASeq and DNA methylation data on breast cancer patients at different stages of the disease. We account for heterogeneity and overdispersion of count data by considering a mixture of negative binomial distributions and incorporate the covariates into the model via a linear modeling construction on the mean components. Our modeling construction employs selection techniques allowing the identification of a small subset of features that best discriminate the samples, simultaneously selecting a set of covariates associated to each feature. Additionally, it incorporates known dependencies into the feature selection process via Markov random field priors. On simulated data, we show how incorporating existing information via the prior model can improve the accuracy of feature selection. In the case study, we incorporate knowledge on relationships among genes via a gene network, extracted from the KEGG database. Our data analysis identifies genes that are discriminatory of cancer stages and simultaneously selects significant associations between those genes and DNA methylation sites. A biological interpretation of our findings reveals several biomarkers that can help to understand the effect of DNA methylation on gene expression transcription across cancer stages.

E0270: **Covariate-dependent graphical models**
Presenter: **Yang Ni**, Texas A&M University, United States
Co-authors: Francesco Stingo, Veerabhadran Baladandayuthapani

Covariate-dependent graphical models are developed. The proposed model allows the graph structure to vary with covariates. We construct conditional independence function that maps from covariate space to sparse graph space. We will present both directed graph and undirected graph cases. Applying the proposed method to multiple myeloma data, we find interesting subject-level gene expression networks.

E0362: **Heterogeneous large datasets integration using Bayesian factor regression**
Presenter: **Alejandra Avalos Pacheco**, Harvard Medical School, Mexico
Co-authors: David Rossell, Richard S Savage
Two key challenges in modern statistical applications are the large amount of information recorded per individual, and that such data are often not collected all at once but in batches. These batch effects can be complex, causing distortions in both mean and variance. We propose a novel sparse latent factor regression model to integrate such heterogeneous data. The model provides a tool for data exploration via dimensionality reduction while correcting for a range of batch effects. We study the use of several sparse priors (local and non-local) to learn the dimension of the latent factors. Our model is fitted in a deterministic fashion by means of an EM algorithm for which we derive closed-form updates, contributing a novel scalable algorithm for non-local priors of interest beyond the immediate scope of this paper. We present several examples, with a focus on bioinformatics applications. Our results show an increase in the accuracy of the dimensionality reduction, with non-local priors substantially improving the reconstruction of factor cardinality, as well as the need to account for batch effects to obtain reliable results. Our model provides a novel approach to latent factor regression that balances sparsity with sensitivity and is highly computationally efficient.

E1345: **Bayesian nonparametric functional mixture modelling to uncover neurocardiovascular profiles in older Irish adults**
Presenter: **Bernardo Nipoti**, Trinity College Dublin, Ireland
Co-authors: Belinda Hernandez
Cardiovascular ageing is one of the principal causes of physical and cognitive disability and mortality. Impaired blood pressure regulation with age is known to influence functional decline. As part of a larger project, which is investigating the relationship between ageing and its links to cardiovascular and neurocardiovascular functioning, we use a Bayesian nonparametric functional mixture model to uncover groups which might identify the major distinct neurocardiovascular profiles in older Irish adults aged 50+. Data were taken from The Irish Longitudinal Study on Ageing (TILDA) and consist of different measurements collected during the so called active-stand experiment.

EO663 Room CLO 101 HIGH DIMENSIONAL TIME SERIES Chair: George Michailidis

E0246: **Monitoring financial networks with online Hurdle models**
Presenter: **Shawn Mankad**, Cornell University, United States
Co-authors: Kamran Paynabar, Mostafa Reisi Gahrooei, Samaneh Ebrahimi
Financial trading data can be used to assess the stability of the financial system. We address this important problem through a network modeling approach that incorporates the complex structure found within high-fidelity trading datastreams. For instance, activities within financial markets can occur in varying market conditions, at different prices, and be of different types and sizes. Furthermore, typically regulators can identify the parties involved in any transaction, allowing for integration with other data, such as information on each counter-party, market announcements, etc. Building on the financial networks literature, which has gained popularity to model these complex dynamics, we create a novel network monitoring system to detect changes within a sequence of sparse networks constructed from an interbank lending market in the European Union. Our approach combines a state space model with the Hurdle model to capture temporal dynamics of the edge formation process, which is modeled as a function of the node and edge attributes and estimated using an extended Kalman Filter. Statistical process control charts are used to monitor the network sequence in real time in order to identify changes in trading patterns that are caused by fundamental shifts in market conditions. We show that the proposed methodology would have raised alarms to the public prior to key events and announcements by the European Central Bank during the 2007-2009 financial crisis.

E1102: **Change point estimation for high-dimensional time series**
Presenter: **Sayar Karmakar**, University of Florida, United States
In many scientific and economic studies, analysis of high-dimensional time series is important. We study high-dimensional sparse VAR models in the context of change-points. We allow the sparse coefficient matrices to evolve over time through multiple change-points. We propose a method to estimate multiple change-points and consequently the coefficient matrices within the successive change-points. Through some recent development in high-dimensional time series we show several theoretical results including consistency of the estimated locations and different pieces. From a practical point of view, since there are humongous number of parameters involved, we also propose a faster way to detect the change-points. Finally, we conclude the paper with some simulations and a real data analysis.

E1253: **Estimation of high-dimensional spectral density matrices**
Presenter: **Sumanta Basu**, Cornell University, United States
Spectral density matrix estimation of multivariate time series is a classical problem in time series and signal processing. In modern neuroscience, spectral density based metrics are commonly used for analyzing functional connectivity among brain regions. We develop a non-asymptotic theory for regularized estimation of high-dimensional spectral density matrices of Gaussian and linear processes using thresholded versions of averaged periodograms. Our theoretical analysis ensures that consistent estimation of spectral density matrix of a p-dimensional time series using samples is possible under high-dimensional regime $\log p = o(n)$, as long as the true spectral density is approximately sparse. A key technical component of our analysis is a new concentration inequality of average periodogram around its expectation, which is of independent interest. Our estimation consistency results complement existing results for shrinkage based estimators of multivariate spectral density, which require no assumption on sparsity but only ensure consistent estimation in a regime $p^2 = o(n)$. In addition, the proposed thresholding based estimators perform consistent and automatic edge selection when learning coherence networks among the components of a multivariate time series. We demonstrate the advantage of our estimators using simulation studies and a real data application on functional connectivity analysis with fMRI data.

E2002: **Asymptotics of large autocovariance matrices**
Presenter: **Monika Bhattacharjee**, IIT BOMBAY, India
High-dimensional moving average process are considered, and the asymptotics for eigenvalues of its sample autocovariance matrices are explored. Under quite weak conditions, we prove, in a unified way, that the limiting spectral distribution (LSD) of any symmetric polynomial in the sample autocovariance matrices, after suitable centering and scaling, exists and is non-degenerate. We use methods from free probability in conjunction with the method of moments to establish our results. In addition, we are able to provide a general description for the limits in terms of some freely independent variables. We also establish asymptotic normality results for the traces of these matrices. We suggest statistical uses of these results in problems such as order determination of high-dimensional MA and AR processes and testing of hypotheses for coefficient matrices of such processes.

EO476 Room CLO 102 CATEGORICAL DATA: ADVANCES AND CHALLENGES Chair: Sabrina Giordano

E0617: **Non-homogeneous interaction effects in the joint action of binary features**
Presenter: **Anna Klimova**, National Center for Tumor Diseases Partner Site Dresden TU Dresden, Germany
Co-authors: Tamas Rudas

Many studies of register data aim to discover association patterns within a set of binary features characterizing the subjects in a population of interest. Unaffected subjects, who possess none of the given features, do not exist in the population, and, therefore, the sample space can be described using an incomplete contingency table where one cell is absent. A new type of interaction parameters which generalize those obtained from the conventional conditional odds ratios is proposed, and, accordingly, a new class of hierarchical log-linear models specified by setting these parameters equal to zero is described. Because the proposed parameters are logarithms of non-homogeneous generalized odds ratios, the resulting models do not include the overall effect (a normalizing constant), and some of their properties are different from those of the classical log-linear models. An example is given to illustrate that the proposed models may allow for describing the association between features in more detail than it can be achieved using a quasi-variant of a classical log-linear model.

E0684: **Robustness of Student link function in multinomial choice models**
Presenter: **Jean Peyhardi**, University of Montpellier, France

The Student distribution has already been used to obtain robust maximum likelihood estimator in the framework of binary choice models. But, until recently, only the logit and probit binary models were extended to the case of multinomial choices, resulting in the multinomial logit (MNL) and the multinomial probit (MNP) models. The recently introduced family of reference models, well defines a multivariate extension of any binary choice model, i.e. for any link function. This paper highlights the robustness of reference models with Student link function, by showing that the influence function is bounded. Inference of the MLE is detailed through the Fishers scoring algorithm, which is appropriated since reference models belong to the family of generalized linear model. These models are compared to the MNL on the benchmark dataset of travel mode choice between Sydney and Melbourne. The results obtained on this dataset with reference models are completely different compared with those usually obtained with MNL, nested logit (NL) or MNP that failed to select relevant attributes. It will be shown that the travel mode choice is totally deterministic according to the terminal waiting time. In fact, the use of Student link functions allow us to detect the total artificial aspect of this famous dataset.

E1150: **Learning procedure for chain (stratified) graphical models**
Presenter: **Manuela Cazzaro**, University of Milano-Bicocca, Italy
Co-authors: Federica Nicolussi, Agnese Maria Di Brisco

The analysis and modeling of multivariate categorical data, usually summarized as contingency tables, is an open issue in statistics. Graphical models are powerful statistical tools to quantify the structure of dependence of these data. Interestingly, the system of (in)dependencies can be represented through a graphical representation such that each vertex of the graph corresponds to a variable and the presence (absence) of an arc between a couple of vertices indicates their functional dependence (independence). Moreover, standard graphical models have been extended to chain stratified graphical models to allow for a greater flexibility in terms of model structure (i.e., permitting marginal, conditional and context-specific independences simultaneously). Several methods exist for Bayesian model determination, i.e. Bayesian learning, of graphical models. In this regard, the first aim is to define a new algorithm of Bayesian learning along the path set out previously. A further aim is to extend the latter algorithm to chain stratified graphical models, thus accounting for context-specific independences. Intensive simulation studies are performed to evaluate the performance of the proposed algorithms and to compare them with existing methods of Bayesian learning.

E1311: **Rating with the pairwise empirical Bayes method**
Presenter: **Cristiano Varin**, Ca Foscari University of Venice, Italy
Co-authors: David Firth

The problem of rating q subjects or items on the basis of a set of n paired comparisons is considered. This problem frequently arises in a variety of fields including psychometrics, analysis of sport tournaments and genetics. Standard analysis of paired comparison data based on Bradley-Terry and Thurstone-Mosteller models becomes difficult in case of sparse tournaments where only a small fraction of all possible paired comparisons is observed. In such situations, empirical Bayes estimation is attractive because it allows borrowing strength across the subjects abilities. Empirical Bayes estimation is numerically impractical in paired comparison models involving a large number of subjects q, because the evaluation of the joint marginal distribution of the data requires approximating intractable q-dimensional integrals. We shall discuss an approach to overcome the numerical difficulties associated with the evaluation of the marginal likelihood through a combination of composite likelihood and empirical Bayes methods.

EO334 Room Court MODELLING AND CLUSTERING COMPLEX DATA II Chair: Geoffrey McLachlan

E1303: **Robust estimation and efficient estimates of partition and model parameters: A step beyond the normality assumption**
Presenter: **Francesca Greselin**, University of Milano Bicocca, Italy
Co-authors: Andrea Cerioli, Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Marco Riani

An iteratively reweighted approach is extended to cluster partitions recently introduced in the literature, and based on multivariate normal clusters, to the case of leptokurtic cores. Assuming multivariate t-distributions for the cluster cores, we estimate their parameters, starting from a high proportion of trimming, and iteratively increasing the active sample size in a controlled fashion. To guarantee consistency at the t-distributed model components, we employ consistency factors to inflate the covariance matrices estimates, based on non-trimmed data. Simulation results and examples on real data show the effectiveness of the approach.

E0887: **Variable screening for high dimensional regression problems via random projections**
Presenter: **Laura Anderlucci**, University of Bologna, Italy
Co-authors: Matteo Farne, Giuliano Galimberti, Angela Montanari

Variable selection in high-dimensional settings characterizes many scientific problems. The concept of sure screening has been previously introduced to reduce the dimensionality, and a procedure has been proposed essentially based on the magnitude of marginal correlations between the predictors and the response variable. We propose a variable selection method for multiple linear regression which is based on axis-aligned random projections and accounts for partial correlation between each predictor and the response. Performances of the proposed method are evaluated on simulated data.

E0904: **Flexible Bayesian modelling of concomitant covariate effects in mixture models**
Presenter: **Giuliano Galimberti**, University of Bologna, Italy
Co-authors: Marco Berrettini, Saverio Ranciati, Thomas Brendan Murphy

Mixtures provide a useful tool to model unobserved heterogeneity and are at the basis of many model-based clustering methods. In order to gain additional flexibility, some model parameters can be expressed as functions of concomitant covariates. In particular, component weights can be linked to concomitant covariates through a multinomial logistic regression model, where each component weight is associated with a linear predictor involving one or more than one concomitant covariate. This approach is extended by replacing the linear predictors with additive ones, where the contributions of some/all concomitant covariates can be represented by smooth functions. In particular, splines are used to approximate these smooth functions. An estimation procedure within the bayesian paradigm is proposed. In particular, a data augmentation scheme based on differenced random utility models is exploited, and smoothness of the covariate effects is controlled by suitable choices for the prior distributions of the spline coefficients. Performances of the proposed methodology are investigated via simulation experiments and some examples on real data are discussed.

E0889: **Matrix sketching in linear discriminant analysis: An efficient strategy for different problems**
Presenter: **Roberta Falcone**, University of Bologna, Italy
Co-authors: Laura Anderlucci, Angela Montanari
Matrix sketching is a recently developed data compression technique. An input matrix A is efficiently approximated with a smaller matrix B, so that B preserves most of the properties of A up to some guaranteed approximation ratio. Thus, numerical operations on big data sets become faster. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. We study the performances of sketching algorithms in the supervised classification context, both in terms of misclassification rate and of boundary approximation, as the degree of sketching increases. We also address, through sketching, the issue of unbalanced classes, which hampers most of the common classification methods.

EO454 Room Jessel ROBUST TESTS FOR CHANGE-POINTS IN TIME SERIES Chair: Herold Dehling

E0890: **Change-point detection based on weighted two-sample U-statistics**
Presenter: **Kata Vuk**, Ruhr-University Bochum, Germany
Co-authors: Herold Dehling, Martin Wendler
A robust change-point test is considered which is based on weighted two-sample U-statistics. We focus on short range dependent data, more precisely on data that can be represented as functionals of a mixing process. In this way, most examples from time series analysis are covered. Under the hypothesis that no change occurs, the limit distribution of the considered test statistic is derived. Under the alternative of a change-point with constant height, we derive consistency. The considered test is sensitive on tails, which means that early and late changes can be better detected. To illustrate the results and to investigate the power of the test, we will give some simulation results.

E1173: **Detecting change-points in time series via ordinal pattern probabilities**
Presenter: **Alexander Schnurr**, University Siegen, Germany
Co-authors: Herold Dehling, Jeannette Woerner, Ines Muenker, Jannis Buchsteiner, Annika Betken
Ordinal patterns describe the order structure of consecutive data points over a small time horizon. Using a moving window approach we reduce the complexity of a time series by analyzing the sequence of ordinal patterns instead of the original data. We present limit theorems for ordinal pattern probabilities and tests for change-points in the short-range dependent as well as in the long-range dependent setting. In the long-range dependent case, we investigate the ordinal information of a subordinated Gaussian process with a non-summable autocovariance function. We establish the asymptotic behavior of different estimators for ordinal pattern probabilities by using a multivariate Hermite decomposition; the limits we obtain (normal -vs- Rosenblatt) depend on the Hermite rank of the functions we consider.

E1822: **Relevant changes in eigensystems of functional time series**
Presenter: **Tim Kutta**, Ruhr-Universitat Bochum, Germany
Co-authors: Holger Dette
Detecting structural changes in time dependent data is a prominent topic in statistical literature. However not all trends in the data are important in applications, but only those of large enough influence. We examine the eigenfunctions and eigenvalues of covariance kernels of $L_2[0,1]$-valued stationary time series for relevant changes. By self normalization techniques we derive pivotal, asymptotically consistent tests for relevant changes and consider their finite sample properties in a simulation study. The investigation of German annual temperature data demonstrates the applicability of our approach.

E0888: **Testing for a change in the tail parameter of regularly varying time series with long memory**
Presenter: **Davide Giraudo**, Ruhr-Universitat Bochum, Germany
Co-authors: Annika Betken, Rafal Kulik
Let $(X_j)_{j\geq 1}$ be a strictly stationary sequence such that the distribution function F of X_1 is regularly varying with parameter $-\alpha$, $\alpha>0$, i.e. $P(X>x)=x^{-\alpha}L(x)$, where L slowly varying. One can show that $\lim_{u\to\infty} E\left(\log(X/u)X>u\right)=\lim_{u\to\infty} E\left(\log(X/u)1_{\{X>u\}}\right)/P(X>u)=1/\alpha=:\gamma$. Consequently, the parameter α can be estimated by $\hat{\gamma}=(1/\sum_{j=1}^n 1_{\{X_j>u_n\}})\sum_{j=1}^n \log(X_j/u_n)1_{\{X_j>u_n\}}$, where $(u_n)_{n\geq 1}$ is a sequence such that $u_n\to\infty$ and $n(1-F(u_n))\to\infty$. To this aim, a two-dimensional empirical process can be used. We will focus on the special case of a volatility model.

EO516 Room MAL 152 SEMI-PARAMETRIC MODELS AND APPLICATIONS Chair: Michael Wiper

E0224: **A Bayesian semi-parametric approach to estimate elliptical distributions**
Presenter: **M Remedios Sillero-Denamiel**, University of Seville, Spain
Co-authors: J Miguel Marin, Pepa Ramirez Cobo, Fabrizio Ruggeri, Michael Wiper
Elliptical distributions are generalisations of the multivariate normal distribution to both longer tails and elliptical contours. We examine how we can use Bayesian methods to perform semi-parametric inference for elliptical distributions. We show how our approach might be used to test whether data are generated from an elliptical distribution or to test whether elliptical data could come from a particular elliptical distribution such as the normal or Student's t.

E0391: **A semi-parametric approach for stochastic frontier models with exogenous variables**
Presenter: **Michael Wiper**, Universidad Carlos III de Madrid, Spain
Co-authors: Yaguo Deng, Helena Veiga
The stochastic frontier model assumes that a firm's output y can be modeled as $y=\beta X-u+v$ where X is a vector of inputs, βX is an efficiency frontier, v is a random error and u is a non-negative, inefficiency term which may depend on exogenous variables. Up to now, most research has assumed a parametric form (exponential, half normal, ...) for modeling the inefficiency distribution, and when exogenous variables are introduced they have typically been included as linear regressors in the location term of this distribution. We address this problem by introducing a semi-parametric, Bayesian model based on Dirichlet process mixtures. Our approach is illustrated with real data applications.

E0431: **Semiparametric nonlinear mixed-effects PK/PD modelling under misspecification**
Presenter: **Dae-Jin Lee**, BCAM - Basque Center for Applied Mathematics, Spain
Mathematical and statistical modelling has become an important tool in the development of new medical drugs, since it may help to understand the outcome of clinical trials. The scientific disciplines concerning mathematical modelling in drug development are called pharmacokinetics and pharmacodynamics (PK/PD). We present an extension of a previous work to address a potential structural model misspecification for population in PK/PD analysis via based on penalized splines. We illustrate our approach with data from a study of the kinetics of the anti-asthmatic drug (theophylline).

E0902: **A semi-parametric approach to bacteria growth modeling**
Presenter: **J Miguel Marin**, University Carlos III, Spain
Co-authors: Michael Wiper, Fabrizio Ruggeri
A new approach is introduced to model multiple bacteria growth curves via the use of a gamma process. The mean of this process can be modeled

using a standard growth curve such as Gompertz or via non-parametric approaches such as splines. The model can also be extended to the case where we observe growth curves under different environmental conditions (temperature, salinity or alkalinity). We show how to fit this model using least squares techniques and illustrate it with an application to Listeria growth curves.

EO318 Room MAL 153 RECENT DEVELOPMENTS IN QUANTILE REGRESSION Chair: Keith Knight

E0791: **Monotone single index models for conditional quantiles**
Presenter: **Roger Koenker**, University of Illinois, United States

Transformation models assert that some linear function of observable covariates when transformed by a univariate function adequately approximates some functional of an observable random response, usually its conditional expectation. When the function is assumed to be monotone, such models can be estimated by profile likelihood; analogous methods for estimating conditional quantile functions will be described and rates of convergence for various target function classes will be discussed.

E1044: **Inference for heterogeneous quantile treatment effects in high dimensions: Rank and score balancing**
Presenter: **Alexander Giessing**, Princeton University, United States

Individuals do not only differ in their innate characteristics, but also in how they respond to a particular treatment or intervention. Quantile regression, which model the effect of covariates on the conditional distribution of the response variable, provide a natural approach to study such heterogeneous treatment effects. We propose a framework for inference of the quantile treatment effect curves in the presence of high-dimensional covariates. Our method combines a de-biased L1-penalized regression adjustment with a quantile-specific balancing scheme which underlies a bias-variance trade-off. By choosing the balancing weights such that they minimize the variance and suppress the bias, we no longer require the covariates distribution to overlap, i.e. the propensity score to be uniformly bounded away from zero and one. We show weak convergence of the proposed estimator to a Gaussian process whose finite dimensional covariance function can be consistently estimated under mild conditions. We discuss the merits of our method and provide finite-sample performances in a variety of settings.

E1287: **Additive quantile regression for electricity demand forecasting**
Presenter: **Matteo Fasiolo**, University of Bristol, United Kingdom

Generalized Additive Models (GAMs) are flexible and interpretable statistical models that are widely used in applied statistics, especially since the advent of efficient and stable methods for smoothing parameter selection. We will describe a computationally efficient and well-calibrated framework for extending GAM methodology to quantile regression models. The proposed methods are based on the general belief updating framework to loss based inference, but they are computed by adapting the stable fitting methods. Fast computation is enabled by adopting a smooth generalization to the quantile regression pinball loss which, if tuned correctly, can also be statistically superior to the original loss. The belief updating framework requires selecting a learning rate balancing the loss with the prior during inference, hence we will present a novel calibration method for selecting this parameter, which aims at obtaining reliable quantile uncertainty estimates. We will briefly discuss the implementation of the proposed methods in the qgam R package, and we will illustrate their performance in the context of an electricity demand forecasting application.

E1377: **Simultaneous estimation and inference for high dimensional censored quantile regression**
Presenter: **Hyokyoung Grace Hong**, Michigan State University, United States

Censored quantile regression has emerged as a powerful tool for detecting heterogeneous effects of covariates on outcomes in survival analysis. With the availability of high dimensional data such as molecular biomarkers, it is often of interest to identify heterogeneous effects of these predictors on patients survival. However, to our knowledge, no work has been conducted to estimate and draw inferences on effects of high dimensional predictors within the framework of censored quantile regression. A novel fused estimator is proposed for modeling survival outcomes with high dimensional predictors and a consistent model-free variance estimator via functional delta method and data re-sampling, which lead to simultaneous statistical inferences for all predictors.

EO272 Room CLO 203 ADVANCES IN DIRECTIONAL STATISTICS Chair: Agnese Panzera

E1264: **Flexible two-piece families of circular distributions**
Presenter: **Jose Ameijeiras-Alonso**, KU Leuven, Belgium
Co-authors: Irene Gijbels, Anneleen Verhasselt

Starting from a base symmetric density and a weight function, a two-piece four parameters density is introduced. The family of unimodal distributions presents a very wide range of skewness and peakedness properties and it allows to generalize some well-known peakedness-free models such as the Batschelet and Papakonstantinou densities. The four parameters of the model have a clear interpretation (mode, concentration, peakedness at the left and at the right of the mode) and symmetric submodels are just obtained when the peakedness parameters are equal. The main properties of the new density are investigated and asymptotic results for maximum likelihood estimators are derived. Finally, the new distribution is applied to real data concerning the time at which the temperature cycle changes.

E1999: **Boxplots for directional data**
Presenter: **Giovanni Camillo Porzio**, University of Cassino and Southern Lazio, Italy
Co-authors: Davide Buttarazzi, Giuseppe Pandolfo, Christophe Ley

Tukey's box-and-whiskers plot is probably one of the most powerful ways to visually provide information on location, variability and symmetries of a univariate distribution. Accordingly, the recently introduced boxplot for circular data will be presented, along to its extension to spherical data. In analogy with the bagplot available for the Euclidean case, a spherical boxplot will be defined. It will be made up of four components: the spherical median, the spherical convex hull of the 50% deepest observations on a sphere (i.e., the box), the spherical convex hull of the remaining points lying within the fences (i.e., the whiskers), the set of far out values (i.e., the potential outliers), if any. The methods will be illustrated through both simulated and real data sets.

E1522: **Tracking the magnetic north pole**
Presenter: **Charles C Taylor**, University of Leeds, United Kingdom
Co-authors: Marco Di Marzio, Agnese Panzera, Stefania Fensore

The problem of forecasting the location of the magnetic north pole is discussed. Based on recent evidence, last movements appear to be not completely explainable by the consolidated knowledge on the subject. Then, it could be desirable to make predictions under very mild assumptions. To this end, we propose a nonparametric approach based on sphere-sphere regression by providing some promising experimental evidence.

E1351: **Some new results on tests for uniformity on the sphere**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium

In dealing with directional data one of the preliminary steps, before any further inference, is to test if such data is isotropic i.e. uniformly distributed around the circle, the sphere or more generally the hypersphere. In view of its importance, there is a considerable literature on the topic. We provide original asymptotic results related to the well-known class of Sobolev tests of uniformity.

EO150 Room CLO 204 RECENT ADVANCES IN NEUROIMAGING STATISTICS Chair: Damian Brzyski

E0318: **Bayesian analysis of fMRI data with spatially-varying autoregressive orders**
Presenter: **Farouk Nathoo**, University of Victoria, Canada
Co-authors: Timothy Johnson, Ming Teng

Statistical modeling of fMRI data is challenging as the data are both spatially and temporally correlated. Spatially, measurements are taken at thousands of contiguous regions, called voxels, and temporally measurements are taken at hundreds of time points at each voxel. Recent advances in Bayesian hierarchical modeling have addressed the challenges of spatiotemproal structure in fMRI data with models incorporating both spatial and temporal priors for signal and noise. While there has been extensive research on modeling the fMRI signal (i.e., the covolution of the experimental design with the functional choice for the hemodynamic response function) and its spatial variability, less attention has been paid to realistic modeling of the temporal dependence that typically exists within the fMRI noise, where a low order autoregressive process is typically adopted. Furthermore, the AR order is held constant across voxels (e.g. AR(1) at each voxel). Motivated by an event related fMRI experiment, we propose a novel hierarchical Bayesian model with automatic selection of the autoregressive orders of the noise process that vary spatially over the brain. With simulation studies we show that our model has improved accuracy and apply it to our motivating example.

E0569: **Bayesian image analysis in transformed spaces**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young

Bayesian image analysis can improve image quality, by balancing a priori expectations of image characteristics, with a model for the noise process via Bayes Theorem. We will give a reformulation of the conventional Bayesian image analysis paradigm in Fourier and wavelet spaces, e.g. for Fourier space the prior and likelihood are given in terms of spatial frequency signals. By specifying the Bayesian model in transformed spaces, spatially correlated priors, that are relatively difficult to model and compute in conventional image space, can be efficiently modeled as a set of independent processes across; the priors are modeled as independent over the transformed space, but tied together by defining a "parameter function" over the space for the values of the pdf parameters. The originally inter-correlated and high-dimensional problem in image space is thereby broken down into a series of (trivially parallelizable) independent one-dimensional problems. We will describe the Bayesian image analysis in transformed space modeling approach, illustrate its computational efficiency and speed, and demonstrate useful properties such as isotropy and resolution invariance to model specification which are difficult to obtain in the conventional formulation. We will describe applications in medical imaging, and contrast with results using conventional Bayesian image analysis models. Finally, we will showcase a Python package that is under development to make the approach widely accessible.

E0987: **Scalable Bayesian Matrix Normal Graphical Models for Brain Functional Networks**
Presenter: **Benjamin Risk**, Emory University, United States
Co-authors: Suprateek Kundu

Recently, there has been an explosive growth in graphical modeling approaches for estimating brain functional networks. In a detailed study, we show that surprisingly, standard graphical modeling approaches for fMRI data may not yield accurate estimates of the brain network due to the inability to suitably account for temporal correlations. We propose a novel Bayesian matrix normal graphical model that jointly models the temporal covariance and the brain network under a separable structure for the covariance to obtain improved estimates. The approach is implemented via an efficient optimization algorithm that computes the maximum-a-posteriori network estimates that is scalable to high dimensions. The proposed method leads to substantial gains in network estimation accuracy compared to standard brain network modeling approaches as illustrated via extensive simulations. We apply the method to resting state fMRI data from the Human Connectome Project (HCP) to study the relationships between fluid intelligence and functional connectivity. Our proposed approach led to the detection of differences in connectivity between high and low fluid intelligence groups, whereas these differences were less pronounced or absent using an existing graphical modeling approach.

E1468: **Bayesian spatial blind source separation via thresholded Gaussian processes**
Presenter: **Jian Kang**, University of Michigan, United States

Blind source separation (BSS) is the separation of latent source signals from their mixtures, which can be achieved by many methods based with different assumptions, criteria or aims, such as principal components analysis (PCA), singular value decomposition (SVD), and independent component analysis (ICA). However, for analysis of neuroimaging data, the most existing BSS methods fail to directly account for the spatial dependence among voxels and do not explicitly model the sparsity of source signals. To address those limitations, we propose a Bayesian nonparametric model for BSS of spatial processes. We assume the observed images as the linear mixtures of multiple sparse and piecewise-smooth latent source processes, for which we construct a new class of prior models by thresholding Gaussian processes. We adopt the von-Mises Fisher distribution as the prior model for mixing coefficients. Under some regularity conditions, we show that the proposed model enjoys large prior support; and we establish the consistency of the posterior distribution with a divergent number of voxels in images. The simulation studies demonstrate that the proposed method outperforms the existing ICA methods for latent brain network separation and brain activation region detection. We apply the proposed method to analysis of the resting-state fMRI data in the Kirby 21 dataset, which can recover existing brain functional networks.

EO592 Room MAL 252 STATISTICS FOR LARGE DIMENSIONAL DATA Chair: Yi He

E0856: **Bayesian model comparison and optimisation**
Presenter: **Liana Jacobi**, University Melbourne, Australia
Co-authors: Dan Zhu, Joshua Chan

The marginal likelihood is the gold standard for Bayesian model comparison. We develop a general framework to automatically compute the sensitivities of marginal likelihood, obtained via simulation-based methods, with respect to any prior hyperparameters, which sets the basis for robustness checks and optimisation routines in Bayesian analysis. Large Bayesian VARs with the natural conjugate prior are now routinely used for forecasting and structural analysis. It has been shown that selecting the prior hyperparameters in a data-driven manner can often substantially improve forecast performance. Using a large US dataset, we show that using the optimal hyperparameter values leads to substantially better forecast performance. Moreover, the proposed method is much faster than the conventional grid-search approach, and is applicable in high-dimensional optimization problems. The new method thus provides a practical and systematic way to develop better shrinkage priors for forecasting in a data-rich environment.

E1015: **Single-index predictive regression**
Presenter: **Hsein Kew**, Monash University, Australia
Co-authors: Weilun Zhou, Jiti Gao, David Harris

A semi-parametric single-index predictive regression model with multiple nonstationary predictors is studied that exhibit co-movement behaviour. Orthogonal series expansion is employed to approximate the unknown link function in the model and the estimator is derived from an optimization under the constraint of identification condition for index parameter. The main finding includes two types of super-consistency rates of the estimators of the index parameter along two orthogonal directions in a new coordinate system. The central limit theorem is established for a plug-in estimator of the unknown link function. In the empirical studies, we provide evidence in favour of nonlinear predictability of the stock return using long term yield and treasury bill rate.

E1027: **Most powerful test against high dimensional alternatives**
Presenter: **Yi He**, University of Amsterdam, Netherlands
Co-authors: Jiti Gao, Sombut Jaidee
A quadratic test in high-dimensional linear regression models is proposed when the number of variables is comparable to, even larger than, the sample size. Using random matrix theory, we show that it has the optimal asymptotic power against non-sparse local alternatives. Our approach allows testing only a subset of coefficients but keeping some nuisance parameters in the estimation, and relaxes the Gaussian assumptions in the literature. Simulation results agree with our asymptotic theory.

E1267: **Robust volatility estimation to semimartingale violations in high-frequency financial data**
Presenter: **Bo Zhou**, Durham University Business School, United Kingdom
Co-authors: Torben Andersen, Yingying Li, Viktor Todorov
The semimartingale assumption for the price process of an asset, traded in a frictionless market, is effectively a no-arbitrage condition. Recently, more and more evidence confirms the existence of violations of the semimartingale property for intraday periods of non-trivial duration. Such violations include, but are not limited to, gradual jumps and bursts in the volatility or drift component. We develop a volatility estimator of integrated volatility (IV), named differenced-increments volatility (DV), robust to these semimartingale violations, while other commonly-used estimators relying on the semimartingale assumption suffer non-trivial finite-sample biases. We document the reliability of our DV estimator in finite samples through a comprehensive Monte Carlo study. In our empirical application, we employ the DV estimator as the predictor to forecast realized volatility (RV). In an application for the S&P 500 index and individual equities, we find that our DV-based Heterogeneous Autoregressive (HAR) model dominates popular competitors according to standard out-of-sample MSE and QLIKE criteria.

EO723 Room MAL 253 ASTROSTATISTICS Chair: Mauro Bernardi

E0300: **Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity**
Presenter: **David Jones**, Texas A&M University, United States
Co-authors: David Stenning, Eric Ford, Robert Wolpert, Tom Loredo, Xavier Dumusque
The radial velocity technique is one of the two main approaches for detecting planets outside our solar system, often referred to as exoplanets. When a planet orbits a star its gravitational force causes the star to move and this induces a Doppler shift (i.e. the star light appears redder or bluer than expected), and it is this effect that the radial velocity method attempts to detect. Unfortunately, these Doppler signals are typically contaminated by various stellar activity phenomena, such as dark spots on the star surface. We propose a Gaussian process modeling framework to capture this stellar activity and thereby improve detection power for low-mass planets (e.g., Earth-like planets). The approach builds on previous work in two ways: (i) we use dimension reduction techniques to construct data-driven stellar activity proxies, as opposed to using standard activity proxies; (ii) we extend a previous multivariate Gaussian process model to a class of models and use a model comparison procedure to select the best model for the particular proxies at hand. Our method results in substantially improved power for planet detection compared with existing methods in the astronomy literature.

E0703: **Astrostatistical challenges in the next decade of cosmology**
Presenter: **Roberto Trotta**, Imperial College London, United Kingdom
Thanks to large and accurate measurements obtained in the last two decades, and to sophisticated statistical analyses, cosmologists have established a cosmological concordance model that reproduces well observations ranging from the relic radiation from the Big Bang to the distribution of galaxies in the sky in the modern Universe. We will introduce and review the observational and theoretical underpinnings of this so-called Lambda-CDM concordance model of cosmology, which strongly points to the existence of both dark matter and dark energy – whose nature is presently the most important outstanding question of cosmology. We will then present the statistical and data science challenges associated with upcoming large observational data from large telescope and space missions. We will discuss how AI and machine learning will be (and already are) indispensable tools to interpret and analyse data about the universe, together with sophisticated Bayesian modeling. We will make the case that we need to go beyond a "black box" approach to inference and model selection, and exploit fully our understanding of the underpinning physics, and our ability to model the data from first principle, including complex selection effects.

E1144: **Continuous time hidden Markov models for astronomical gamma-ray light curves**
Presenter: **Andrea Sottosanti**, University of Padua, Italy
Co-authors: Mauro Bernardi, Alessandra Rosalba Brazzale, Luis Campos, Aneta Siemiginowska, David van Dyk
The detection and characterisation of celestial objects is an inherently inter-disciplinary field which embraces both statistical and astronomical methods. Pioneering technology has driven remarkable acceleration in the rate of detection of celestial objects, and global space astrometry missions will produce accurate maps by surveying stars at an ever-increasing rate. This rapid progression through technology has determined a paradigm shift in observational astronomy, where large digital sky surveys are becoming the dominant source of data. At the same time, astrostatistics has rapidly evolved during the past two decades into a stand-alone discipline with its own professional associations. Indeed, the simple collection of rich, massive data sets on the terabyte scale is not the end of the process but just its beginning. Beyond the technological improvements, a key step in astronomical breakthrough research is the meaningful statistical analysis of the collected information. We introduce a novel approach for the analysis and characterisation of the light curves emitted by time-varying high-energy astronomical phenomena based on continuous time hidden Markov models. The proposed method analyses the variation of signal from a source in time and successfully identifies different latent states that correspond to distinct physical mechanisms. We provide also a bootstrap procedure to evaluate the nature of extreme values in the light curve.

E1241: **Detecting new signals under background mismodelling**
Presenter: **Sara Algeri**, University of Minnesota, United States
When searching for new astrophysical phenomena, uncertainty arising from background mismodelling can dramatically compromise the sensitivity of the experiment under study. Specifically, overestimating the background distribution in the signal region increases the chance of missing new physics. Conversely, underestimating the background outside the signal region leads to an artificially enhanced sensitivity and a higher likelihood of claiming false discoveries. The aim is to provide a unified statistical algorithm to perform modelling, estimation, inference and signal characterization under background mismodelling. The proposed method allows us to incorporate the (partial) scientific knowledge available on the background distribution, and provides a data-updated version of it in a purely nonparametric fashion, without requiring the specification of prior distributions. If a calibration sample or control regions are available, the solution discussed does not require the specification of a model for the signal; however, if the signal distribution is known, it allows further improvement of the accuracy of the analysis and detection of additional signals of unexpected new sources.

EG127 Room MAL G16 CONTRIBUTIONS IN COPULAS AND DEPENDENCE MODELLING Chair: Takeshi Emura

E1643: **Smooth bootstrapping of copula functionals**
Presenter: **Klaus Herrmann**, Sherbrooke University, Canada
Co-authors: Maximilian Coblenz, Oliver Grothe, Marius Hofert
The smoothed bootstrap for functionals defined on the set (or possibly only a subset) of copulas is considered. Examples for such functionals

include measures of association such as Kendall's tau or Spearmans' rho, the upper and lower tail dependence coefficients, or level sets that are used to quantify the risk inherent in joint events. The investigation is motivated by the question of how much the smoothing aspect of smoothed bootstrapping influences the underlying dependence structure in a multivariate framework. The strength of this dependence distortion may depend on the functional, the smoothing kernel or the sample size. We address these points with a special focus on elliptical distributions and smoothing kernels. While most motivating examples are bivariate in nature, the discussion is valid in arbitrary dimensions, making the results viable for high-dimensional settings and data science applications in general. A crucial part of multivariate kernel estimation, and hence of our algorithm, is the selection of a suitable bandwidth matrix. While the literature on bandwidth selection for multivariate kernel distribution function estimation has so far focused on special cases such as product kernels or diagonal bandwidth matrices, we present a novel cross-validation based approach that is valid for general bandwidth matrices.

E1827: **Modelling the association in bivariate survival data by using a Bernstein copula**
Presenter: **Mirza Nazmul Hasan**, Hasselt University, Belgium
Co-authors: Roel Braekers

Bivariate or multivariate survival data arise when a sample consists of clusters of two or more subjects which are correlated. The focus is on clustered bivariate survival data that are possibly censored. Two approaches are commonly used in modelling such type of correlated data: random effects models and marginal models. A random effects model includes a frailty model and assumes that subjects are independent within a cluster conditionally on a common non-negative random variable, the so-called frailty. In contrast, the marginal approach models the marginal distribution directly and then imposes a dependency structure through copula functions. Bernstein copulas are used to account for the correlation in modelling bivariate survival data. A two-stage parametric estimation method is developed to estimate in the first stage the parameters in the marginal models and in the second stage the coefficients of the Bernstein polynomials in the association. Hereby we use a penalty parameter to make the fit desirably smooth. In this aspect, linear constraints are introduced to ensure uniform univariate margins, and we use quadratic programming to fit the model. We perform a simulation study and illustrate the method on a real data set.

E1839: **EM algorithms for estimating the B-spline copula**
Presenter: **Xiaoling Dou**, Waseda University, Japan
Co-authors: Satoshi Kuriki

The B-spline copula is a generalization of the Bernstein copula. It is defined by replacing the Bernstein basis functions by B-spline basis functions. This change requires the copula parameters satisfy slightly different conditions, in spite of the copula form remains the same. Because the Bernstein copula can be considered as a finite mixture distribution for given marginals, we can use EM algorithm methods to estimate the Bernstein copula. Since this idea is also available for the B-spline copula, we propose to generate the existing EM algorithms of the Bernstein copula to estimate the B-spline copula by changing the basis functions and the parameter conditions. Illustrative examples are presented with real data sets.

E1675: **Copulas for multiple time series: Evidence of asymmetric price transmission along the Italian pork supply chain**
Presenter: **Giorgia Rivieccio**, Parthenope University, Italy
Co-authors: Giovanni De Luca, Fabian Capitanio, Barry Goodwin

The presence of asymmetric vertical price contagions in the Italian pork market is investigated, describing the multiple dependence structure along the supply chain and evaluating the degree of extreme value dependence at different market levels. We provide a nonlinear copula-based Vector AutoRegressive (VAR) model built under a three stage estimation method. We contribute to the literature, providing a specification of the copula-based Impulse Response Functions (IRF), giving also an interpretation of the economic results to the empirical analysis.

EG113 Room Senate CONTRIBUTIONS IN EXTREME VALUES Chair: Armelle Guillou

E1734: **Parameter estimation of the generalized Pareto distribution for normal baseline distribution**
Presenter: **Eva Lopez Sanjuan**, Universidad de Extremadura, Spain
Co-authors: M Isabel Parra Arevalo, Jacinto Martin Jimenez, Mario Martinez Pizarro

In Extreme Value Theory, the estimation of the limit distribution is usually made discarding an important amount of data. When we apply peaks-over-threshold method, for Generalized Pareto Distribution (GPD), only values above a certain threshold are considered and much information is wasted. We employ all the available data to make estimations for the parameters of the GPD, taking advantage of the existing relationship between the parameters of baseline distribution and the limit ones. We focus on the case when the baseline distribution is Normal. Different simulations were carried out in order to compare the effectiveness of this strategy to the standard Metropolis-Hastings algorithm. Besides, the accuracy of the method is tested employing real data, provided by Red Automatica de Monitoreo Atmosferico (RAMA), corresponding to the levels of PM2.5, the most dangerous polluting agent in Ciudad de Mexico between 2003 and 2019.

E1737: **An improved prior choice for Gumbel distribution parameters to model extreme values**
Presenter: **M Isabel Parra Arevalo**, Universidad de Extremadura, Spain
Co-authors: Eva Lopez Sanjuan, Mario Martinez Pizarro, Francisco Javier Acero Diaz

The Gumbel distribution can be employed to model the maximum (or the minimum) of a sequence of observations. Bayesian estimation (block maxima method) for its two parameters is usually performed by using only record values of the observations, consequently a lot of information is wasted. The proposed method seizes all the available observations in order to increase the accuracy of the estimations. The key is to consider the existing relationship between the parameters of the baseline distribution for the observations and the ones for the extreme Gumbel distribution. To evaluate the performance of the proposed method, replicated data sets from different baseline distributions in the Gumbel domain of attraction are simulated. Overall, the results show that the proposed method outperforms block maxima method for the estimation of both parameters.

E1775: **Asymptotic comparison of second-order parameters estimators**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal
Co-authors: Ivanilda Cabral, Frederico Caeiro

Under an adequate third-order framework, and for heavy right tails, i.e., a positive extreme value index, an asymptotic comparison of alternative estimators of shape and scale second-order parameters is developed. Such an asymptotic comparison is performed for a fixed k, the number of upper order statistics involved in the estimation, and at optimal levels, i.e., at levels k_0 where the asymptotic mean square error of the estimators is minimized.

E1622: **Functional covariate-adjusted extremal dependence**
Presenter: **Anwar Albdulathem**, University of Edinburgh, United Kingdom

A method is proposed that tracks how the dependence between the extreme values of a random vector may change conditionally on a random function. Our model can be regarded as a functional covariate regression model, tailored for situations where there is the need of assessing how extremal dependence changes according to a random function. The main target of interest is what we define as the angular manifold, which is a family of angular densities indexed by a functional covariate. The methods are motivated by the need of evaluating how the dependence between extreme losses in two stock markets (e.g. NYSE and NASDAQ) changes according to the shape of a certain random curve (e.g. Daily Treasury Yield Curve). To estimate the family of angular densities on the angular manifold, we follow a similar line of attack as a popular approach for extending the Nadaraya-Watson estimator to the functional context. Our estimator can be regarded as a version of a previous one, and the simulation

study suggests that the proposed methods perform well in wealth of simulation scenarios.

EG579 Room MAL 251 CONTRIBUTIONS IN ENVIRONMENTAL APPLICATIONS Chair: Serge Guillas

E1719: **Estimating multispecies biodiversity indicators using hidden Markov models**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece

Biodiversity indicators play a vital role in the assessment of changes in biodiversity. Several indicators have been recently developed, based on derived indices of abundance for individual species. We describe a state-space model formulation for the calculation of multispecies biodiversity indicators. The new approach has several advantages, including the ability to deal with unavailable data when species enter late in the study, incorporate estimates of uncertainty of the species index values when these are available, and allow smoothing within the indicator, rather than posthoc, as is frequently the case. We show how hidden Markov modelling may be used to efficiently fit the model in general, and illustrate the approach using abundance data from a 'citizen science' scheme on butterflies in the UK.

E1881: **Analyzing the relationship between seismic spectrum, water discharge and bedload transport in dynamic gravel-bed rivers**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Jordi Diaz Cusi, Javier Alvarez Pulgar, Gil Gonzalez-Rodriguez, Elena Fernandez Iglesias, Daniel Vazquez Tarrio, Jorge Marquinez

Seismic data are typically employed to monitor earthquake activity, but they can also be exploited in order to investigate the existing links between the seismic signal and a broad range of physical processes occurring in the nearby rivers. For instance, the noise related to water turbulence during high discharges has a clear impact on the seismic signal. Concurrently, high-flows in gravel-bed rivers involve the displacement of large volumes of coarse sediment particles, travelling as bedload, which in turn should be the source of an additional and non-negligible amount of seismic noise. Data derived from the 3-years seismic monitoring of a gravel-bed river located in NW of Spain are considered. First, we have accomplished spectral analysis over the seismic noise generated by more than 10 river flow events with high discharge. Then, we have applied several statistical analysis not only for calibrating the functions describing the links between seismic noise and fluvial discharge but also in order to identify the thresholds for detection of incipient bedload transport and geomorphic change in the seismic signal.

E0351: **Parametric post-processing of dual resolution precipitation forecasts**
Presenter: **Marianna Szabo**, University of Debrecen, Hungary
Co-authors: Sandor Baran

In recent years, all major weather prediction centres issue ensemble forecasts which are obtained from multiple runs of numerical weather prediction models with various initial conditions and model parameterizations. The European Centre for Medium-Range Weather Forecasts (ECMWF) produces operational ensemble-based analyses and predictions that describe the range of possible scenarios and their likelihood of occurrence. According to its strategic plans till 2025, ECMWF wants to improve the resolution of ensemble forecasts, which requires a substantial increase of computation resources. Researchers at the ECMWF experiment with a mixture of high and low resolution ensemble forecasts to determine the optimal combination on a fix computational cost. We investigate the effect of statistical post-processing of gridded ECMWF dual resolution ensemble precipitation forecast for Europe. We apply the censored shifted gamma distribution based semi-local ensemble model output statistics as our approach for calibration.

E1594: **Heteroscedastic autoregressive postprocessing of temperature forecasts**
Presenter: **Annette Moeller**, Clausthal University of Technology, Germany
Co-authors: Juergen Gross

To account for uncertainty in numerical weather prediction (NWP) models it has become common practice to employ ensembles of NWP forecasts. However, forecast ensembles often exhibit forecast biases and dispersion errors, thus require statistical postprocessing to improve reliability of the ensemble forecasts. An extension of a recently developed postprocessing model utilizing autoregressive information present in the forecast error of the raw ensemble members is proposed. The original approach is modified to let the variance parameter additionally depend on the ensemble spread, yielding a two-fold heteroscedastic model. Furthermore, a high-resolution forecast is included into the postprocessing model, yielding improved predictive performance. Finally, it is outlined how the autoregressive model can be utilized to postprocess ensemble forecasts with higher forecast horizons, without the necessity of making fundamental changes to the original model. To illustrate the performance of the heteroscedastic extension of the autoregressive model, and its use for higher forecast horizons we present a case study for a data set containing 12 years of temperature forecasts and observations over Germany. The case study indicates that the autoregressive model yields particularly strong improvements for forecast horizons beyond 24 hours ahead.

CO212 Room Bloomsbury PREDICTIVE MODELLING AND TIME SERIES Chair: Jonas Andersson

C1616: **Fraud detection by a multinomial model: Separating honesty from unobserved fraud**
Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway
Co-authors: Andreas Olden, Aija Rusina

With the problem of detecting tax evasion in mind, we investigate how to identify items, e.g. individuals or companies, that are wrongly classified as honest. Normally, we observe two groups of items, labeled fraudulent and honest, but suspect that many of the observationally honest items are, in fact, fraudulent. The items observed as honest are therefore divided into two unobserved groups, honestH, representing the truly honest, and honestF, representing the items that are observed as honest, but that are actually fraudulent. By using a multinomial logit model and assuming commonality between the observed fraudulent and the unobserved honestF, a method that uses the EM-algorithm to separate them has been previously presented. By means of a Monte Carlo study, we investigate how well the method performs, and under what circumstances. We then compare it to other standard methods.

C1573: **Online high dimensional covariance change point detection**
Presenter: **Clifford Lam**, London School of Economics and Political Science, United Kingdom

Detecting changes in the covariance structure in time series data is an important problem which finds applications in, e.g., clustering, identification of disease related genes in bioinformatics, or portfolio and risk management in finance, to name but a few areas. In high dimensional time series data where the dimension can be larger than the sample size, this problem can be extremely difficult. For one, the typical statistic for change detection involves the calculation of sample covariance matrix of two different parts of data, which can be affected by the poor performance of sample covariance matrix under high dimensional setting. More importantly, as far as we know there are no covariance detection methods so far that can have very imbalanced sample sizes for the two parts of data involved, which unfortunately is exactly the setting for online covariance detection, where one part can have good number of data points, but the newer part of the data may have only finite number of data points. We propose a series of statistics which can be powerful in covariance change detection under the online setting, with the dimension of each observation vector grows together with or even faster than the sample size, while the newer part of the data under suspicion of change can have only finite sample size. Asymptotic normality of these statistics under both no change and change scenarios are proved and demonstrated with numerical examples.

C1654: **Spatial models with smooth transitions**
Presenter: **Ingrid Mattsson**, Uppsala university, Sweden

Co-authors: Johan Lyhagen

The aim is to investigate spatial autoregressive models and the possibility that spatial effects differ between different locations. To account for this we propose a model where a logistic function is used to capture the non-linearity in the spatial lag parameter. This creates a spatial lag model with smooth transitions, where the effect of the spatial neighbors depends on the transition variable in the logistic function. An LM test for detecting non-linearity is derived and a simulation study, where the properties of the test are investigated, is conducted. The simulations reveal that the test shows good power even in relatively small samples with moderate deviation from linearity. We further include an empirical application where data from the 2014 Swedish general election is used to explore the spatial dependence between voting districts.

C1708: **A smooth transition duration model with an application to the deregulation of the Queensland electricity market**

Presenter: **Paulina Joneus**, Department of Statistics, Uppsala University, Sweden

Co-authors: Johan Lyhagen

The smooth transition duration model is introduced. It is designed to model the dependence of duration on explanatory variables allowing for a structural change. The proposed model is a generalization of parametric survival regression models and makes it possible to detect a non-linear behaviour when the response of interest is the time until some event occurs. A Lagrange multiplier test of the null hypothesis of linearity is derived together with the maximum likelihood estimators of the smooth transition duration model. The impact of the deregulation in the Queensland electricity market on the electricity spot price is then assessed by examining the time between abnormal price increases. The deregulation might have led to a change in the behaviour of the market participants and the smooth transition duration model is used to detect and examine such an eventual transition. The preliminary results show a clear support to a gradual change in the appearance of abnormal price increases.

CO406 Room G11 NETWORK ECONOMETRICS AND FINANCIAL NETWORKS Chair: Monica Billio

C0253: **Oil shocks and production network structure: Evidence from the OECD**

Presenter: **Petre Caraiani**, Bucharest University of Economic Studies; Institute for Economic Forecasting, Romania

Using a Bayesian time-varying VAR, the impact of oil shocks on GDP for a set of the OECD economies is derived. Various measures are further estimated to characterize the production network structure based on Input-Output matrices. When we analyze the relationship between the time-varying responses of GDP to oil demand and oil supply shocks and production network characteristics, we find that measures like skewness in the in-degrees and in the out-degrees or density tend to amplify the negative impact of oil shocks on GDP. The results are in line with the recent literature that outlines the importance of network structures for aggregate dynamics.

C0850: **A Bayesian graphical VAR model for yield curve fluctuations**

Presenter: **Monica Billio**, University of Venice, Italy

Co-authors: Andrea Berardi, Roberto Casarin

Yield curve fluctuations across different currency areas are generally highly interrelated. However, both the contemporaneous causal relationships and the temporal dependence structure vary over time. We document the time-varying behaviour of the degree of connectedness among yield changes in seven currency areas (Australia, Canada, Germany, Japan, Switzerland, UK and US). We decompose yields into expected short rates and term premia using a Gaussian ATSM integrated with long-term yield expectations and analyse the contribution of those components to global yield co-movements and connectedness. We find that the dependence structure of both yields and their components can be significantly different for short and long maturities. The empirical analysis is based on a Bayesian graphical VAR model, where the contemporaneous and temporal causal structures of the structural VAR are represented by two different graphs and an efficient Markov chain Monte Carlo algorithm is used to estimate jointly the two causal structures and the parameters of the reduced-form VAR model. When representing bond yields as a network, the increased system fragility is reflected by a degree distribution which is symmetric and has thinner tails, whereas asymmetry and fat tails suggest that there is heterogeneity in the linkages among countries.

C1247: **Looking at Bayesian nonparametric clustering from a community detection point of view**

Presenter: **Stefano Tonellato**, Ca' Foscari University of Venice, Italy

It is well-known that a wide class of Bayesian nonparametric priors leads to the representation of the distribution of the observables as a mixture density with an infinite number of components, and that such a representation induces a clustering structure in the observations. However, cluster identification is not straightforward a posteriori and some post-processing of the MCMC output is usually required. It has been proven that pairwise posterior similarity successfully allows to either apply classical clustering algorithms or estimate the underlying partition by minimising a suitable loss function. We show how it can be used to map sample items on a weighted undirected graph, where each node represents an individual and edge weights are given by the posterior pairwise similarities. A community detection algorithm, known as infomap, can be applied to such a network, providing a minimum description length unsupervised classification. A relevant feature of this method is that it allows for the quantification of the posterior uncertainty of the clustering. The same approach can be easily extended to the unsupervised classification of time series.

C1813: **A network centrality approach to stock returns: Sectoral interdependences, communities, and volatility transmission**

Presenter: **Sebastiano Michele Zema**, Scuola Superiore Sant'Anna di Pisa, Italy

Simple sectoral influence strength measures are provided for equity returns which yield a hierarchical taxonomy between the sectors of the economy. Covering mega, large, and midcap public companies stably listed in the US during the period 2000-2018, financial companies are found to be central in the network and the main contributors to the market mode. The stocks' tendency to cluster in communities is thus explored by means of the Louvain algorithm over different periods and for different thresholds imposed on the correlation matrix, showing the topological collapse of the stock market during the crises. The identified communities follow the GICS classification partially and only out of turbulent periods. The hierarchical structure detected in the stock market at the micro level is exploited, showing how few stocks, topologically central, are statistically significant in a Granger causality framework for the realized volatility of the aggregate, while stocks in the periphery do not possess any forecast power. This few confirm the rules behavior to be robust under different network metrics, providing useful insights regarding the role played by market actors in shaping aggregate volatility.

CO586 Room G21A ADVANCES IN REALIZED VOLATILITY ESTIMATION Chair: Ekaterina Kazak

C0435: **BUMVU estimators**

Presenter: **Aleksey Kolokolov**, Manchester Business School, United Kingdom

Co-authors: Roberto Reno, Patrick Zoi

The theory of Uniformly Minimum Variance Unbiased Estimators is extended to the special class of Block estimators. We provide necessary and sufficient conditions for a block estimator to have uniformly minimum variance. We show the relevance of this theory uncovering new results in two classical statistical problems: estimation of the error term variance in homoskedastic nonparametric regressions, and estimation of volatility functionals of semimartingales observed at a high frequency.

C0704: **Renewal based volatility estimation**

Presenter: **Yifan Li**, The University of Manchester, United Kingdom

Co-authors: Ingmar Nolte, Sandra Nolte

The aim is to develop the idea of renewal time sampling, a novel sampling scheme constructed from stopping times of semimartingales. Based

on this new sampling scheme we propose a class of volatility estimators named renewal based volatility estimators. We show that: (1) The spot variance of a continuous martingale can be expressed in terms of the conditional intensity or conditional duration density of renewal sampling times; (2) In an infill asymptotics setting, renewal based volatility estimators are consistent and jump-robust estimators of the integrated variance of a general semimartingale; (3) Renewal time sampling and range-based sampling have a higher sampling efficiency than equidistant return-based sampling.

C0863: **Predicting financial risk with artificial neural networks: Whether there is information in high frequency returns**
Presenter: **Christian Muecher**, University of Konstanz, Germany

High Frequency financial data is vastly used for modelling financial risk, often by utilizing Realized Variance estimators. This aim is to model financial risk directly using high frequency returns as inputs for an artificial neural network. Artificial neural networks are universal approximators and thus are able to learn the function that predicts the conditional variance. There exist approaches showing that the function learned by a neural network using past daily returns as an input and the next periods squared daily return as an output is a consistent estimator of the conditional variance function. We extend these approaches by directly using high frequency returns as input to learn the variance function. The results using this more extensive information are compared to the approach based on past daily returns, as well as existing, standard, approaches used for predicting the variance of financial assets. The comparison is done in terms of simulated price processes and an application to real data.

C0692: **Portfolio choice: Balancing forecasting risk**
Presenter: **Ekaterina Kazak**, University of Manchetser, United Kingdom
Co-authors: Ingmar Nolte, Sandra Nolte, Yifan Li

Estimation noise is a well-known issue in empirical portfolio modelling. Estimated weights are known to have huge standard errors and bad predictive quality, which often results in an inferior out-of-sample portfolio performance compared to simple alternatives. Most of the recent literature concentrates on the improvement of covariance matrix forecasts, which would hopefully result in better portfolio performance. However, the proposed models often suffer from the dimensionality problem, such that the forecasting error still dominates the theoretical gain. We propose a portfolio choice model, which explicitly takes into account forecasting risk and avoids the dimensionality problem by forecasting a one-dimensional portfolio measure directly. We then define a forecasting error based on the realized measures and look for weight estimates which results in the more precise forecast in terms of the forecasting error variance and at the same time is not far from the optimal portfolio solution. The proposed approach is close to the James-Stein type of estimator, which balances bias-variance trade-off in a data-driven manner. The proposed method is shown to outperform the commonly used approaches in both simulation and empirical studies.

CO240 Room G4 FORECASTING IN FINANCIAL MARKETS Chair: Robinson Kruse-Becher

C0525: **IVX-based panel predictive regressions for stock returns**
Presenter: **Christoph Hanck**, Universitat Duisburg-Essen, Germany
Co-authors: Matei Demetrescu

New panel tests are proposed for predictability of, e.g., stock returns through regressors which are allowed to be persistent or stationary. The panel units may be heterogeneous and cross-sectionally dependent. Building on previous work, we employ IVX instruments to this end, which are generated within the panel and require no outside exogeneity assumptions. We show the test statistics to follow standard χ^2 distributions under the null of no predictability as the number of time series observations T and panel units N jointly go to infinity. Simulations indicate good size and power in panels of size commonly encountered in empirical practice.

C0526: **Robust dynamic portfolio choice based on out-of-sample performance**
Presenter: **Rainer Alexander Schuessler**, University of Rostock, Germany

A new approach is introduced to solve dynamic portfolio choice problems with a focus on robust out-of-sample performance. We therefore devise a strategy that rigorously tackles the problem of estimation error. The method involves defining a discrete set of single-period portfolio allocation policies (candidate portfolio strategies) and choosing among them at portfolio revision dates, relying on bootstrapped pseudo out-of-sample portfolio returns. A key aspect of the approach involves providing candidate portfolio strategies that generate (approximately) iid portfolio returns. We apply the method to dynamic investment problems in futures trading, strategic asset allocation and a cross-sectional momentum strategy in equity markets.

C0550: **A dynamic functional factor model for yield curves: Identification, estimation, and prediction**
Presenter: **Sven Otto**, University of Bonn, Germany

The problem of yield curve forecasting from a functional time series perspective is discussed. A functional factor model is considered, in which the factors follow some linear autoregressive process. The model is identified by imposing suitable conditions on the factors and the loading functions. By applying the least squares principle, a functional principal components based estimator is obtained, which is shown to be consistent. The minimum mean squared error forecast from the dynamic functional factor model is considered, and pointwise and simultaneous prediction bands are derived. Finally, the accuracy of the predictions and prediction bands is discussed in an out-of-sample experiment with monthly yield curves of U.S. Treasuries.

C0520: **A robust evaluation of macro-financial predictive content for realized volatility**
Presenter: **Robinson Kruse-Becher**, University of Cologne, Germany

Predictive regressions for realized volatility are studied. Dating back to 1989, there is an ongoing debate whether financial and macroeconomic series have predictive power for financial volatility. We make use of the econometric recently provided framework to study the role of several potential predictors for realized volatility in a robust way. In contrast to standard approaches, the applied methodology accounts for long memory, multiple structural breaks (in levels and persistence) and spurious regressions. As most predictors are highly persistent, and realized volatility has typical long memory features, it is important to account for these phenomena when testing for predictive power. Standard predictive regressions method fail in this context and might even provide spurious evidence in favor of predictability. We employ an updated monthly data set used previously which covers a long time span and several important predictors. Among these are credit spreads, term spreads, price-dividend and price-earnings ratios from 1885 to 2016. Our findings indicate important differences in the outcomes when properly accounting for changes in persistence and long memory. We further study the role of time-variation in the predictive content by recursive estimation and test for spurious long memory in a number of robustness checks.

CO562 Room Gordon RECENT ADVANCES IN QUANTILE REGRESSION Chair: Carlos Lamarche

C1035: **Exact computation of censored least absolute deviations estimator**
Presenter: **Yannis Bilias**, Athens University of Economics and Business / RC, Greece

Quantile Regression (QR) in the presence of censoring results in objective functions that need to be optimized which are non-convex and non-smooth. Approximate optimization algorithms proposed for the practitioners do not guarantee the finding of global optimizer. Under this scenario, the statistical properties of the QR estimator are not known and its use will lead to invalid inference. We propose the use of modern optimization methods for locating the global optimum in this class of estimation problems. We address the exact computation of Censored Least Absolute Deviations (CLAD) estimator by formulating the estimator as a linear Mixed Integer Programming (MIP) problem with disjunctive constraints.

Application of our approach to previously studied datasets suggests that widely used approximate optimization algorithms can lead to erroneous conclusions. The exact computation of global optimum also allows us to compare the statistical properties of the Powell's estimator with those of other asymptotically equivalent competitors that are easier to compute.

C1045: **Inference for shape constrained quantile regression splines**
Presenter: **Thomas Parker**, University of Waterloo, Canada

Inference methods are proposed for nonparametric quantile regression estimates based on series methods that are subject to shape constraints such as that the conditional quantile curves are monotone or convex in an explanatory variable. Constraints can also be imposed to maintain monotonicity in quantile level. Hypotheses such as linearity of conditional quantile curves against the alternative of convex curves are considered across quantile levels. Inference is nonstandard but can be conducted using resampling.

C1082: **High-dimensional predictive quantile regression with mixed roots**
Presenter: **Rui Fan**, Rensselaer Polytechnic Institute, United States
Co-authors: Ji Hyung Lee

The benefit of using adaptive LASSO for predictive quantile regression is studied. The commonly used predictors in predictive quantile regression typically have various degrees of persistence, and exhibit different signal strength in explaining the conditional quantiles of dependent variable. We show that the adaptive LASSO methods have the consistent variable selection and the oracle properties under the simultaneous presence of stationary, unit root and cointegrated predictors. Some encouraging simulation results are reported.

C1208: **Inference for penalized quantile regression for panel data**
Presenter: **Carlos Lamarche**, University of Kentucky, United States
Co-authors: Thomas Parker

Penalized quantile regression is a relatively new technique for estimation of panel quantile models. The existing literature has been mostly focused on the consistency of the point estimator under different assumptions. We investigate inference in a class of penalized quantile regression estimators based on wild bootstrap procedures. Simulation studies are carried out to investigate the small sample behavior of the proposed approaches. Finally, we illustrate the application of the new approaches using a real data example.

CO408 Room Montague ADVANCES IN CREDIT RISK MODELLING I Chair: Raffaella Calabrese

C0702: **Modeling exposure at default under varying systematic conditions**
Presenter: **Maximilian Nagl**, University of Regensburg, Germany
Co-authors: Daniel Roesch, Jennifer Betz

In the advanced internal ratings based approach, banks are allowed to use own estimates of exposure of default to determine their regulatory capital. For volatile segments, downturn estimates i.e., estimates which reflect economic downturn conditions are demanded. Furthermore, banks are obliged to base their models on credit conversion factors for credit lines. This regulatory setting might be challenging. First, the distribution of credit conversion factors is highly bimodal and reminds of loss rate distributions. Second, downturn estimates require an adequate consideration of systematic effects which might be non trivial. A unique data set of defaulted credit lines from the U.S. and Europe is used, and a quantile regression approach is applied to model conditional distributions of credit conversion factors. If macroeconomic variables are incapable of revealing the true systematic patterns, the model is enhanced by random effects to ensure adequate downturn estimates. To the best of the authors' knowledge, this is the first attempt to provide a sound framework to model the full conditional distribution of credit conversion factors which ensures adequate downturn estimates considering observable and unobservable systematic variables.

C1177: **Joint models for longitudinal and survival data with INLA: Applications in credit scoring**
Presenter: **Victor Medina-Olivares**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Jonathan Crook

Joint models for longitudinal and survival data are an appealing modelling framework for credit scoring since they allow to jointly model the time to default and the internal time-varying covariates usually seen in credit risk data. However, the estimation procedure is computationally expensive and sometimes unfeasible for the size of the data in the credit context. Although, if the joint model is assumed with a linear bivariate Gaussian association structure, then it can be seen as a latent Gaussian model (LGM) and thus the Bayesian inference can be approximated with the integrated nested Laplace approximation (INLA). We propose a joint model for longitudinal and survival data in a discrete-time setting with applications in credit scoring and estimated with INLA.

C1206: **Modelling spatial contagion effects for mortgage defaults using a Bayesian hierarchical approach**
Presenter: **Matteo Spada**, Paul Scherrer Institute, Switzerland
Co-authors: Raffaella Calabrese

The aim is to explore how the spatial location of US mortgage loans can improve the predictive accuracy of credit risk models. The recent financial crisis led to much concern about the so-called credit contagion - how the deterioration of a borrowers future ability to honour its debt obligations can affect the ability of other borrowers living in the same neighbourhood. Consequently, the spatial contagion effects is suggested to be included in the risk assessment for mortgage loans. A Bayesian hierarchical model based on the geographical locations of properties is proposed to measure the impact of the credit contagion on the overall risk. This is a robust and fully probabilistic approach that incorporates spatial correlation and enables to model both the observed data and any unknowns as random variables. Such a method is applied to loan-level data released by Freddie Mac on US single-family mortgages.

C1683: **Default probability models applied to a Mexican peasant institution**
Presenter: **Maria Rosa Nieto Delfin**, Investigaciones y Estudios Superiores SC, Mexico
Co-authors: Jose Morfin Tarasco

Credit risk models failure to forecast crises has become especially important since the global economic crisis of 2008. The mortgage backed assets that started the crisis had excellent credit ratings. Since then, global credit risk regulators, are converging to have more efficient regulations which aim to develop accurate credit risk models. It is postulated that the Value at Risk forecast of a peasant financial company is more accurate if an endogenous stochastic model is used to calculate the default probability. A variety of models were tested on the historic data of the peasant financial company to find the best fit one. The results show that a Zero-Adjusted-Inverse-Gaussian model is the best fit for this type of credit institution. Hence, the Value at Risk forecast of the peasant financial company is improved. It was also found that Mexican credit risk regulations are undesirable, as they prohibit institutions from calculating their credit risk parameters through internal models. Mexican regulators give a generic value for the default probability vector to each institution. If the current credit risk regulations in Mexico changed to allow institutions to calculate their risk parameter through internal models, they would improve the calculation of its Value at Risk.

CO432 Room Woburn APPLIED MACROECONOMIC AND MACRO-FINANCIAL TOPICS I Chair: Christoph Gortz

C0291: **Is there news in inventories?**
Presenter: **Christoph Gortz**, University of Birmingham, United Kingdom
Co-authors: Christopher Gunn, Thomas Lubik

Inventories are an important, highly volatile and forward looking component of the business cycle, yet they have been largely neglected by the literature on TFP news shocks that argues such shocks are important drivers of macroeconomic fluctuations. We use a standard VAR identification to document a new fact: in response to TFP news, inventories move procyclically along with the other major macroeconomic aggregates. Our finding is not self-evident: conventional views would suggest news about higher future productivity provides incentives to run the current inventory stock down and increase stockholding in the future when productivity is high. We provide evidence that this substitution effect is dominated by a demand effect due to which firms increase inventories in response to sales in light of rising consumption and investment. Our empirical fact corroborates the view that TFP news shocks are important drivers of macroeconomic fluctuations. However, it imposes a challenge to existing theoretical frameworks as they fail to reproduce the procyclical inventory movements in response to TFP news shocks. We suggest this comovement puzzle can be solved through extending a standard framework with intangible capital.

C1252: **Identification, informational sufficiency and the role of monetary policy**
Presenter: **Haroon Mumtaz**, Queen Mary University of London, United Kingdom
Co-authors: Mirela Sorina Miescu

Informational sufficiency and a valid identification strategy are both necessary conditions to recover reliable impulse responses in structural vector autoregressive models (SVAR). We recommend the use of a Proxy Factor Augmented VAR model (FAVAR) which addresses the two conditions in a unified framework. We evaluate the performance of the Proxy FAVAR model versus a small scale Proxy SVAR in two Monte Carlo experiments. We show that the Proxy FAVAR model outperforms the Proxy SVAR in several cases of miss-specification as well as when the instrument is contaminated. In an empirical exercise, we examine the effects of monetary policy shocks on a large set of variables. We find that the impulse responses from a Proxy FAVAR model are considerably different from the ones delivered by a Proxy VAR, especially for real activity and prices variables. The results suggest that for an accurate evaluation of the effects of monetary policy shocks, it is crucial to complement a valid identification strategy with a large information set.

C1316: **Monetary policy and wealth inequality over the great recession in the UK: An empirical analysis**
Presenter: **Angeliki Theophilopoulou**, Brunel university, United Kingdom
Co-authors: Haroon Mumtaz

The UK has experienced a dramatic increase in wealth and income inequality over the past four decades. By using detailed micro information at household level from the Wealth and Assets Survey, we construct monthly historical measures of wealth inequality from 2005 to 2016. We investigate the dynamic relationship between conventional and unconventional monetary policy and whether it played a role in the evolution of wealth inequality measures. The findings suggest that expansionary monetary policy shocks lead to an increase in wealth inequality in the UK and contribute significantly to its fluctuation. The heterogeneous response of wealth at different quantiles suggests that financial easing has a larger positive effect on high income households and the portfolio channel had a stronger impact than the labour income channel to the low percentiles of the distribution. Our evidence also suggests that the policy of quantitative easing may have contributed to the increase in inequality over the Great Recession.

C1478: **Macroeconomic effects from expectations of ECB's asset purchases**
Presenter: **Stephane Lhuissier**, Banque de France, France

The macroeconomic effects of anticipations about future net asset purchases of the European Central Bank (ECB) are examined. To do so, we first construct a new indicator of markets' expectations about the final size of ECB's asset purchases using on Bloomberg and Reuters surveys, conducted ahead each ECB Governing Council. We show that markets had, most of the time, anticipated the official announcements of ECB's asset purchases. Second, we use this indicator in a VAR Framework estimated on monthly data from November 2014 to June 2019. We show that an asset purchase anticipation of one percent of GDP leads to a rise of 0.14 percent in real GDP and 0.08 percent of prices. Third, we run several counterfactual simulations to establish the role of markets' expectations in shaping business cycle fluctuations.

CO576 Room Chancellor's Hall THEORY AND APPLICATION OF PREDICTIVE REGRESSIONS Chair: Robert Taylor

C0421: **Testing in predictive quantile regressions with time-varying volatility**
Presenter: **Matei Demetrescu**, CAU Kiel, Germany
Co-authors: Robert Taylor, Paulo Rodrigues

While stock return predictability has received considerable attention in the literature, predictability tests are geared at detecting whether the conditional mean of the return series of interest depends on putative predictors or not. To help decide on quantile predictability, we discuss tests based on the Langrange Multiplier principle. The LM approach leads to a simple auxiliary regression, for which inference can be conducted using instrumental variable estimation. We therefore obtain simple linear IV-based tests that are robust to conditional and unconditional heteroskedasticity. Moreover, we provide an analysis of the behavior of the proposed tests under a factor structure of the regressors, where the common and idiosyncratic components may be either near-integrated or stable autoregressions or both. We find the proposed tests to perform well in finite samples compared with alternatives based on quantile regressions and resampling.

C0400: **Testing for episodic predictability in stock returns**
Presenter: **Robert Taylor**, University of Essex, United Kingdom
Co-authors: Paulo Rodrigues, Matei Demetrescu, Iliyan Georgiev

Standard tests based on predictive regressions estimated over the full available sample data have tended to find little evidence of predictability in returns. Recent approaches based on subsamples have been considered, suggesting that predictability might exist only within so-called 'pockets of predictability'. These methods are prone to the criticism that the sub-sample dates are endogenously determined. To avoid this we propose new tests based on the maximum of statistics from sequences of forward and backward recursive, rolling, and double-recursive predictive sub-sample regressions. We show that the limiting distributions of our proposed tests are robust to both the degree of persistence and endogeneity of the regressors in the predictive regression, but not to any heteroskedasticity present even if the sub-sample statistics are based on heteroskedasticity-robust standard errors. We therefore develop fixed regressor wild bootstrap implementations which we demonstrate to be first-order asymptotically valid. Finite sample behaviour against a variety of temporarily predictable processes is considered. An empirical application to US stock returns illustrates the usefulness of the new predictability testing methods we propose.

C0601: **Performance of predictive regressions when models are uncertain**
Presenter: **Benjamin Hillmann**, Kiel University, Germany

In spite of well-developed theory, predictability of stock returns using fundamental variables is difficult to establish in practice. One possible explanation for this fact is potential nonlinearity of the predictive relation; another possible explanation is of a rather statistical nature, namely the uncertainty associated with estimation, and in particular model selection. The effectivity of various forecasting procedures is explored taking these aspects into account. In an out-of-sample forecasting exercise, proper accounting for nonlinearity and model uncertainty is shown to somewhat improve evidence on predictability.

C1081: **Long-run predictability revisited**
Presenter: **Paulo Rodrigues**, Universidade Nova de Lisboa, Portugal

Co-authors: Matei Demetrescu, Robert Taylor

Long-run predictability is revisited and a simple yet powerful new approach is proposed. The procedure is based on the bias reduced IVX framework recently proposed which extends previous contributions. The new approach has several advantages over existing procedures designed to test for long-run predictability. The first is its simplicity of application, which makes it very appealing for empirical applications, when compared, for instance, with Bonferroni based methods, as these require the computations of confidence intervals for the near-integrated parameter c which characterises the persistence of the predictor; the second advantage is that left, right and two sided hypotheses tests can be immediately computed from our approach, whereas Bonferroni based methods require the computations of different confidence intervals for c depending on whether left or right tailed intervals are to be computed; and third our approach is easily generalized to a multi-predictor context.

Saturday 14.12.2019 14:35 - 16:15 Parallel Session E – CFE-CMStatistics

EI008 Room Beveridge Hall DIRECTIONS IN STATISTICAL MODELLING Chair: Jochen Einbeck

E0164: **Investigating the latent structure of criminal networks**
Presenter: **Isabella Gollini**, University College Dublin, Ireland

A new latent variable modelling approach is presented to investigate the latent structure of criminal networks. This allows us to explain the relational structure of the data by estimating the positions of the suspects in a latent social space. In particular, we illustrate this new methodology by exploring a complex network consisting of interdependent ego-networks based on the wiretaps acquired by the Italian Police in 2014 on 29 suspects (egos) during an investigation about human smuggling out of Libya. The statistical challenge with these ego-networks is that the large number of alters (more than 15k) can potentially be members of several ego-networks. Moreover, from a computational point of view, this model is difficult to estimate due to the intractability of the likelihood. To efficiently overcome this difficulty we adopt an efficient variational algorithm. The flexible modelling framework introduced can be adapted to a wide range of network settings.

E0165: **On the use of clustering in a predictive model**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Christophe Biernacki, Mohammed Sedki, Vincent Vandewalle

Many data, in biostatistics, contain some sets of variables which permit evaluating unobserved traits of the subjects (e.g., we ask question about how many pizzas, hamburgers, chips... are eaten to know how healthy are the food habits of the subjects). Moreover, we often want to measure the relations between these unobserved traits and some target variables (e.g., obesity). Thus, a two-steps procedure is often used: first, a clustering of the observations is performed on the sets of variables related to the same topic; second, the predictive model is fitted by plugging the estimated partitions as covariates. Generally, the estimated partitions are not exactly equal to the true ones. We investigate the impact of these measurement errors on the estimators of the regression parameters, and we explain when this two-steps procedure is consistent. We also present a specific EM algorithm which simultaneously estimates the parameters of the clustering and predictive models.

E0606: **On whether to check the model before doing model-based inference**
Presenter: **Christian Hennig**, University of Bologna, Italy
Co-authors: Iqbal Shamsudheen

Statistical inference comes with model assumptions, and it is a standard recommendation to check the model assumptions before doing model-based inference. A problem with this is that checking model assumptions affects subsequent inference. Even in case a model assumption is in fact fulfilled, it is no longer fulfilled conditionally on passing a model misspecification test (misspecification paradox). In the literature there is some scattered investigation of how big a problem this is, and whether the resulting combined procedures (i.e., choosing the inference method depending on whether certain assumptions are passed or not) are advisable. Much of this work is surprisingly critical of such a practice. Several aspects of such combined procedures are discussed and new results are presented, investigating theoretically and by simulation setups in which fulfilled and violated model assumptions are mixed. This provides a more positive if still not uncritical view of such combined procedures.

EO574 Room CLO B01 FUNCTIONAL SHAPE DATA ANALYSIS Chair: Sonja Greven

E0613: **Shape-motivated functional data analysis**
Presenter: **Anuj Srivastava**, Florida State University, United States

Functional data has a growing presence in all branches of science and engineering, partly due to tremendous advances made in data collection and storage technologies. Such data is mostly analyzed using the classical Hilbert structure of square-integrable function spaces, but that setup ignored shapes of functions and leads to counter intuitive results. Shape implies the ordering and the heights of peaks and valleys but is flexible on their exact locations. To focus on shapes of functions, we have introduced Elastic functional data analysis that allows time warpings of functions in order to register functional data, i.e. match their peaks and valleys. This, in turn, requires elastic Riemannian metrics that enable comparisons and testing of shape data modulo warping group action. We will present some statistical tools resulting from this framework, including estimation of shape-constrained functions and probability densities.

E0652: **Visualization and outlier detection for multivariate elastic curve data**
Presenter: **Sebastian Kurtek**, The Ohio State University, United States
Co-authors: Weiyi Xie, Oksana Chkrebtii

A new method is proposed for the construction and visualization of geometrically-motivated boxplot displays for elastic curve data. We use a recent shape analysis framework, based on the square-root velocity function representation of curves, to extract different sources of variability from elastic curves, which include location, scale, shape, orientation and parametrization. We then focus on constructing separate displays for these various components using the Riemannian geometry of their representation spaces. This involves computation of a median, two quartiles, and two extremes based on geometric considerations. The outlyingness of an elastic curve is also defined separately based on each of the five components. We evaluate the proposed methods using multiple simulations, and then focus our attention on real data applications. In particular, we study variability in (a) 3D spirals, (b) handwritten signatures, (c) 3D fibers from diffusion tensor magnetic resonance imaging, and (d) trajectories of the Lorenz system.

E1233: **Elastic analysis of irregularly and sparsely sampled curves**
Presenter: **Lisa Steyer**, Humboldt University of Berlin, Germany
Co-authors: Almond Stoecker, Sonja Greven

Even though functional shape data is assumed to consist of continuous curves (up to invariances), these curves are usually not observed in practice. In real applications, the outline of an object, for example a bone, is often observed only at a small number of discrete points, which even might not correspond for different curves. To approximate the elastic distance between irregularly and sparsely sampled curves, we interpret them as polygons, hence treat them as piece-wise linear. We can show that the warping problem simplifies in this case where at least one of the curves is piece-wise linear, and use this to improve computations. We use this approximation to provide distance-based methods for observed curves modulo warping and to compute smooth means for samples of curves.

E1237: **Component-wise gradient boosting for functional shape regression**
Presenter: **Almond Stoecker**, Humboldt University of Berlin, Germany
Co-authors: Sonja Greven

In 2D, the shape of an object, such as a bone or a body cell, may be represented either by a collection of prominent points (landmark shape) or by its outline (functional shape), which are considered modulo the shape preserving transformations of translation, rotation and scaling. Component-wise gradient boosting for regression with shape responses is developed allowing for modular specification of multiple linear or smooth covariate effects in an additive predictor. Boosting is particularly attractive in this context due to its step-wise procedure and inherent regularization, which provides automated variable selection and is thus well suited for complex model scenarios potentially involving many covariates.

EO598 Room G11 RECENT ADVANCEMENTS IN CAUSAL INFERENCE Chair: Joseph Antonelli

E0211: **Patterns of effects and sensitivity analysis for differences-in-differences**
Presenter: **Luke Keele**, University of Pennsylvania, United States

Applied analysts often use the differences-in-differences (DID) method to estimate the causal effect of policy interventions with observational data. The method is widely used, as the required before and after comparison of a treated and control group is commonly encountered in practice. DID removes bias from unobserved time-invariant confounders. While DID removes bias from time-invariant confounders, bias from time-varying confounders may be present. Hence, like any observational comparison, DID studies remain susceptible to bias from hidden confounders. We develop a method of sensitivity analysis that allows investigators to quantify the amount of bias necessary to change a study's conclusions. Our method operates within a matched design that removes bias from observed baseline covariates. We develop methods for both binary and continuous outcomes. We then apply our methods to two different empirical examples from the social sciences. In the first application, we study the effect of changes to disability payments in Germany. In the second, we re-examine whether election day registration increased turnout in Wisconsin.

E1119: **Measurement error-robust causal inference via synthetic instrumental variables**
Presenter: **Caleb Miles**, Columbia University, United States
Co-authors: Brent Coull, Linda Valeri

While measurement error is known to be benign in certain settings, this is often not the case when estimating causal effects. Two scenarios in which it can be malignant are the estimation of (i) the average causal effect when confounders are measured with error and (ii) the natural indirect effect when the exposure and/or confounders are measured with error. Methods adjusting for measurement error typically require external data or knowledge about the measurement error distribution. We propose methodology not requiring any such information. Instead, we show that when the outcome regression is linear in the error-prone variables, consistent estimation of these causal effects can be recovered using what we refer to as synthetic instrumental variables. These are functions of only the observed data that behave like instrumental variables for the error-prone variables. Using data from a study in Bangladesh, we apply our methodology to estimate (i) the effect of maternal protein intake on child neurodevelopment while controlling for lead exposure, and (ii) maternal protein intake's role in mediating the effect of lead exposure on child neurodevelopment. Protein intake is calculated from food journal entries, and is suspected to be highly subject to measurement error.

E1239: **The analysis of relative treatment effects in multi-drug-resistant tuberculosis from fused observational studies**
Presenter: **Mireille Schnitzer**, Universite de Montreal, Canada
Co-authors: Guanbo Wang, Arman Alam Siddique, Asma Bahamyirou, Dick Menzies, Andrea Benedetti

Multi-drug-resistant tuberculosis (MDR-TB) is defined as strains of tuberculosis that do not respond to at least the two most used anti-TB drugs. After diagnosis, the intensive treatment phase for MDR-TB involves taking several alternative antibiotics concurrently. The Collaborative Group for Meta-analysis of Individual Patient Data in MDR-TB has assembled a large, fused dataset of over 30 observational studies comparing the effectiveness of 15 antibiotics. The particular challenges that we have considered in the analysis of this dataset are the large number of potential drug regimens, the resistance of MDR-TB strains to specific antibiotics, and the identifiability of a generalized parameter of interest though most drugs were not observed in all studies. We describe causal inference theory and methodology that we have applied or developed for the estimation of treatment importance and relative effectiveness of different antibiotic regimens with a particular emphasis on targeted learning approaches.

E1309: **Propensity score regression and G-estimation**
Presenter: **David Stephens**, McGill University, Canada

The links between g-estimation and regression adjustment are reviewed by using the propensity score. G-estimation is well established as a doubly robust and locally efficient semiparametric inference procedure under certain assumptions, but has not been as widely used as other adjustment procedures perhaps due to it being less transparent in its construction, and more difficult to implement. The binary outcome case has proved particularly challenging for g-estimation, although procedures have now been developed for this case. We will show that using a regression construction inference in the binary case can be implemented very straightforwardly using standard tools, and that the regression-based procedure constructs the semiparametric efficient estimator. Finally, we will demonstrate the use of g-estimation in the context of dynamic treatment rule construction utilizing a semiparametric model selection criterion.

EO360 Room G3 RISK MEASURES, INFERENCE, AND APPLICATIONS Chair: Ricardas Zitikis

E0374: **Estimation of risk measures**
Presenter: **Thorsten Schmidt**, University Freiburg, Germany

While risk measures are a topic of paramount importance, the estimation of risk measures has long been neglected in the literature. Most of the practically used estimation procedures introduce a bias in the sense that the risk is underestimated. This is confirmed in backtesting procedures, where the performance shows potential of improvement. We present unbiased estimators which do not suffer from this deficiency. Moreover, we present a new estimation procedure based on deep neural networks which allows us to obtain unbiased estimators in a numerically efficient way when explicit formulae are not available.

E0364: **Pension fund ALM with multivariate second order stochastic dominance constraints**
Presenter: **Sebastiano Vitali**, University of Bergamo, Italy
Co-authors: Milos Kopa, Vittorio Moriggia

A pension fund manager typically decides the allocation of the pension fund assets looking for a long-term sustainability. Many Asset and Liability Management models in the form of multistage stochastic programming problem have been proposed to help the pension fund manager to define the optimal allocation given a multi-objective function. The recent literature proposes multivariate stochastic dominance constraints to guarantee that the optimal strategy is able to stochastically dominate a benchmark portfolio in a multistage framework. We extend previous results to a new type of multivariate stochastic dominance. In particular, instead of considering multiple single-stage stochastic dominance constraints or a linear combination of the stages, we apply a unique constraint that involves jointly multiple stages. Numerical results show the difference between the different ways to interpret and apply the multivariate stochastic dominance.

E0644: **Explaining the risk behind futures prices with distortion functions**
Presenter: **Daniela Escobar**, London School of Economics, United Kingdom
Co-authors: Florentina Paraschiv, Michael Schuerle

It is well-known that the classical arbitrage-based methods are not applicable for pricing electricity derivatives in the same way as for other commodities since electricity cannot be stored. Consequently, futures valuation do not follow a general rule. We fill the gap between spot prices and futures prices with three different premia: the distortion premium, an information premium and an ambiguity premium. Firstly, distortion functions serve to explain the different risk preferences of consumers and producers when trading these futures. Besides, we make the distortion premium applicable for pricing electricity futures, by including negative risk premia. Secondly, it has been established that futures prices contain more information than the one provided by the spot market. Therefore, we include a correction parameter to quantify this lack of information and define an information premium. Lastly, we include an ambiguity premium which is measured in terms of the Wasserstein distance. The goal is to identify these three different components from observed prices and shed some light on the pricing mechanism of futures and their risk premia. For

our results, we propose a regime-switching model for the spot prices. Overall, this methodology allows the identification of distortion functions and a correction factor under model ambiguity.

E0367: **An axiomatic foundation for the expected shortfall**
Presenter: **Ricardas Zitikis**, University of Western Ontario, Canada
Co-authors: Ruodu Wang

The Value-at-Risk (VaR) and the Expected Shortfall (ES) are the most popular risk measures used in banking and insurance regulation. According to the recent Basel Accords, ES has replaced VaR as the standard risk measure for market risk in the banking sector. VaR has been characterized in the literature with several sets of economic axioms, whereas ES, although being a most popular and coherent risk measure, does not yet have an axiomatic foundation. We shall put forward four intuitively attractive economic axioms that uniquely characterize ES. Key to the characterization are novel notions such as p-tail event and p-concentration, and we shall discuss them in detail.

EO072 Room G4 ROBUST MODELLING Chair: Eva Cantoni

E1323: **Tuning and smoothing parameter selection for robust estimation of non-parametric effects**
Presenter: **William Aeberhard**, Stevens Institute of Technology, United States
Co-authors: Eva Cantoni, Rosalba Radice, Giampiero Marra

Deviations from model assumptions are known to throw off any likelihood-based estimation and inference, and hinder penalization schemes meant to ensure some degree of smoothness for non-parametric (additive, non-linear) effects approximated by linear combinations of basis functions. Robust estimation methods are a reliable alternative, but the usual tuning of an efficiency-robustness trade off based on asymptotic covariances is not meaningful any more since the achieved smoothness is generally not comparable between estimation methods. To address this, we propose a median downweighting proportion criterion which is simple, general, and fast to compute. Furthermore, we extend the Fellner-Schall smoothing parameter selection method to robust estimation and compare it to robust versions of the Akaike Information Criterion and Schwarz's Bayesian Information Criterion. Various generalized additive models are used for illustration.

E0778: **A robust version of GAMLSS**
Presenter: **Rosalba Radice**, Cass Business School, United Kingdom
Co-authors: William Aeberhard, Eva Cantoni, Giampiero Marra

A robust version of generalised additive models for location, scale and shape is discussed where any parameter of the distribution can be specified as function of additive predictors allowing for several types of covariate effects (e.g., linear, non-linear, random and spatial effects). The estimation approach permits all models parameters to be estimated robustly by limiting the influence of deviating data points on each log-likelihood contribution. We evaluate the empirical performance of the proposed method through simulation experiments. We also illustrate the use of this approach on functional magnetic resonance imaging measurements for a human brain subject to a particular experimental stimulus.

E0688: **Robustness and confidence distributions**
Presenter: **Laura Ventura**, University of Padova, Italy
Co-authors: Erlis Ruli, Monica Musio

Inferential topics for a parameter of interest (such as reaching point estimates, assessing their precision, setting up tests along with measures of evidence, finding confidence intervals, comparing the value of the parameter of interest with other parameters from other studies, etc.) may be automatically performed if a frequentist distribution, without prior, is available. An approach to derive a frequentist distribution is based on confidence distributions (CDs) and confidence curves. A CD analysis is much more informative than providing a confidence interval or a p-value. The standard theory for parametric inference evolves around the use of likelihood methods, and this is also partly the case for CDs. Typically, to first-order, CDs may be based on the large sample theory for the maximum likelihood estimator, the Wald statistic and the likelihood-ratio test. The basic concepts and recipes for CDs are however not limited to likelihood methods, and various alternatives may be worked with. For instance, it is well known that in the presence of model misspecifications, likelihood methods may be inaccurate. The aim is to discuss the use of robust unbiased estimating equations in order to compute a robust CD. In particular, we suggest both asymptotic robust CDs obtained by using first-order results for estimating equation inference and a simulation-based approach to CD, based on a frequentist reinterpretation of Approximate Bayesian Computation techniques.

E1375: **Privacy-preserving parametric inference: A case for robust statistics**
Presenter: **Marco Avella-Medina**, Columbia University, United States

Differential privacy is a cryptographically-motivated approach to privacy that has become a very active field of research over the last decade in theoretical computer science and machine learning. In this paradigm one assumes there is a trusted curator who holds the data of individuals in a database and the goal of privacy is to simultaneously protect individual data while allowing the release of global characteristics of the database. In this setting we introduce a general framework for parametric inference with differential privacy guarantees. We first obtain differentially private estimators based on bounded influence M-estimators by leveraging their gross-error sensitivity in the calibration of a noise term added to them in order to ensure privacy. We then show how a similar construction can also be applied to construct differentially private test statistics analogous to the Wald, score and likelihood ratio tests. We provide statistical guarantees for all our proposals via an asymptotic analysis. An interesting consequence of our results is to further clarify the connection between differential privacy and robust statistics. In particular, we demonstrate that differential privacy is a weaker stability requirement than infinitesimal robustness, and show that robust M-estimators can be easily randomized in order to guarantee both differential privacy and robustness towards the presence of contaminated data. We illustrate our results both on simulated and real data.

EO775 Room G5 NEW METHODS AND MODELS FOR TIME SERIES ANALYSIS Chair: Carmela Cappelli

E1186: **Studying the influence of economic sectors on stocks through a partial dependence analysis**
Presenter: **Giovanni De Luca**, University of Naples Parthenope, Italy
Co-authors: Marta Nai Ruscone, Giorgia Rivieccio

Understanding the complex nature of financial markets is still a great challenge. In particular, a challenge is to understand the underlying mechanisms of influence that operate in financial markets. A method is discussed to estimate how a company is influenced by an economic sectors after identifying the partial dependence structure of each asset with the assets of the sector excluding the influence of the market. An effective way used to capture the dependence structure of a multivariate time-series is the copula function. However the variety of copula functions and ease of estimation dramatically reduce when the dimension of the multivariate time-series increases. On the other hand, bivariate copula functions are popular and effective in capturing the dependence structure of a 2-dimensional continuous random vector. A simple strategy to continue to use bivariate copula functions for modelling multivariate time-series is the recourse to the vine copulas. The procedure provides a picture of the relative influence of the economic sectors on each stock in terms of Kendall's τ and tail dependence.

E1515: **A theoretical regression tree for classifying risky financial institutions**
Presenter: **Carmela Cappelli**, University of Naples Federico II, Italy
Co-authors: Francesca Di Iorio, Angela Maddaloni, Pierpaolo Durso

A recursive partitioning approach to identify groups of risky financial institutions is proposed by using a synthetic indicator built on the information arising from a sample of pooled systemic risk measures. The composition and amplitude of the risky groups changes over time, emphasizing the periods of high systemic risk stress. We also calculate the probability that a financial institution can change risk group over the next month and show that a firm belonging to the lowest or highest risk group has in general a high probability to remain in that group.

E1697: **Bootstrap prediction intervals for weighted SETAR forecasts**
Presenter: **Marcella Niglio**, University of Salerno, Italy
Co-authors: Francesco Giordano

The generation of prediction intervals is not always an easy task, mainly when the interest is focused on nonlinear time series models whose predictor distribution is not standard. After presenting a new predictor for the Self Exciting Threshold AutoRegressive (SETAR) model, we focus the attention on the generation of the prediction interval. In more detail, the new predictor, that we call weighted SETAR predictor, is obtained as a weighted mean of the past observations. The weights are obtained from the minimization of the Mean Square Forecast Errors (MSFE). We propose a residual bootstrap procedure to build prediction intervals that are then compared to other approaches largely used in the literature though a Monte Carlo study. In particular, the coverage of the different prediction intervals is examined, considering SETAR models with increasing degree of complexity and nonlinearity.

E1929: **Unit roots in periodic time series**
Presenter: **Domenico Cucina**, Roma Tre University, Italy
Co-authors: Francesco Battaglia, Roberto Baragona

Many time series are subject to seasonal fluctuations that might not be constant over time. It has been shown that such series may be well described by Periodic AutoRegressive (PAR) models, in which each season of the year follows a possibly different AR process. When seasonality also contains a stochastic component, the problem of seasonal unit roots, associated with changing seasonality, arises. The study of seasonal unit roots in periodic autoregressive models merits attention because it allows a more complete description of the seasonal component. There exist many tests to examine if a monthly time series has seasonal unit roots for autoregressive processes. The so-called HEGY method, for example, tests the presence of unit root and considers different combinations of constant, trend and seasonal dummies. We propose an empirical method to examine whether a periodic model has one or more seasonal unit roots, and to detect them.

EO426 Room Gordon STATISTICAL METHODS FOR TIME-VARYING MULTIVARIATE DATA Chair: Irina Gaynanova

E0749: **A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data**
Presenter: **Ricardo Monti**, Gatsby Computational Neuroscience Unit, UCL, United Kingdom

Connectivity estimation is challenging in the context of high-dimensional data. A useful preprocessing step is to group variables into clusters, however, it is not always clear how to do so from the perspective of connectivity estimation. Another practical challenge is that we may have data from multiple related classes (e.g., multiple subjects or conditions) and wish to incorporate constraints on the similarities across classes. We propose a probabilistic model which simultaneously performs both a grouping of variables (i.e., detecting community structure) and estimation of connectivities between the groups which correspond to latent variables. The model is essentially a factor analysis model where the factors are allowed to have arbitrary correlations, while the factor loading matrix is constrained to express a community structure. The model can be applied on multiple classes so that the connectivities can be different between the classes, while the community structure is the same for all classes. We propose an efficient estimation algorithm based on score matching, and prove the identifiability of the model. Finally, we present an extension to directed (causal) connectivities over latent variables. The practical utility of method is demonstrated on two resting-state fMRI datasets: the first corresponds to data from the Autism Brain Imaging Data Exchange (ABIDE) consortium and the second to data from the Cambridge Center for Ageing and Neuroscience (CamCAN) repository.

E1281: **Change point estimation in a dynamic stochastic block model**
Presenter: **George Michailidis**, University of Florida, United States

The problem of estimating the location of a single change point in a dynamic stochastic block model is considered. We propose two methods of estimating the change point, together with the model parameters. The first employs a least squares criterion function and takes into consideration the full structure of the stochastic block model and is evaluated at each point in time. Hence, as an intermediate step, it requires estimating the community structure based on a clustering algorithm at every time point. The second method comprises the following two steps: in the first one, a least squares function is used and evaluated at each time point, but ignores the community structures and just considers a random graph generating mechanism exhibiting a change point. Once the change point is identified, in the second step, all network data before and after it are used together with a clustering algorithm to obtain the corresponding community structures and subsequently estimate the generating stochastic block model parameters. A comparison between these two methods is illustrated. Further, for both methods under their respective identifiability and certain additional regularity conditions, we establish rates of convergence and derive the asymptotic distributions of the change point estimators. The results are illustrated on synthetic and real data.

E1321: **Large scale maximum average power multiple inference on time-course count data with application to RNA-Seq analysis**
Presenter: **Jay Breidt**, Colorado State University, United States
Co-authors: Wen Zhou, Meng Cao, Graham Peers

Experiments that longitudinally collect RNA sequencing (RNA-seq) data can reveal dynamic patterns of differential gene expression. Most existing tests are designed to distinguish among conditions based on overall differential patterns across time, though in practice, a variety of composite hypotheses are of more scientific interest. Further, existing methods may lack power and some fail to control the false discovery rate (FDR). We propose a new model and testing procedure to address these issues simultaneously. Conditional on a latent Gaussian mixture with evolving means, we model the data by negative binomial distributions, introduce a general testing framework based on the proposed model and show that the proposed test enjoys the optimality property of maximum average power. The test allows not only identification of traditional differentially-expressed genes, but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and demonstrate its good performance via simulation studies. The procedure reveals interesting biological insights when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of a marine diatom.

E1334: **Time-varying canonical correlation analysis**
Presenter: **Grace Yoon**, Texas A and M University, United States
Co-authors: Irina Gaynanova

Canonical correlation analysis (CCA) has been widely used to describe associations between two sets of variables and multiple extensions have been developed for high-dimensional data. However, existing methods cannot be applied to data over time. We present time-varying CCA which can be applied to high-dimensional data and allows us to study how the associations between the two data sets change over time. We also propose data aggregation based on the change point for the small sample size. In addition, the proposed method can take into account variable types of two data sets, such as continuous, zero-inflated and binary. We evaluate the performance of the proposed method in both simulated and real data.

EO276 Room MAL G13 ADVANCES IN BAYESIAN METHODS Chair: David Rossell

E0373: **Partially factorized and tighter variational Bayes for probit models**
Presenter: **Daniele Durante**, Bocconi University, Italy
Co-authors: Augusto Fasano
Bayesian regression models for dichotomous data arise in several applications. Within such a framework, recent research has shown that the posterior distribution for the p probit coefficients has a unified skew-normal kernel, under Gaussian priors, and hence can be expressed via a convolution of p-variate Gaussians and n-variate truncated normals with full covariance matrix. Such a novel result allows efficient Bayesian inference for a wide class of applications, but closed-form calculation of posterior moments and predictive distributions is unfeasible for large sample sizes due to the intractability of multivariate truncated normals with dependent components. To address this issue we propose a variational approximation for the unified skew-normal posterior which factorizes the multivariate truncated Gaussian component via a product of univariate truncated normals. We prove that such a result can be formally interpreted as a partially factorized mean-field variational Bayes strategy which provides a tighter approximation to the posterior distribution for the probit coefficients, compared to state-of-the-art solutions, while crucially preserving skewness. A simple coordinate ascent variational inference algorithm is developed and the improved performance is outlined in simulations and applications.

E0498: **Criteria for Bayesian hypothesis testing in two-sample problems**
Presenter: **Victor Pena**, Baruch College, City University of New York, United States
Two criteria for prior choice in two-sample testing are proposed which have a common starting point: a hypothetical situation where perfect knowledge about one of the groups is attained, while the data for the other group are assumed to be fixed. In such a scenario, the Bayes decision of the two-sample problem should arguably converge to the Bayes decision of a one-sample test where the distribution of the group for which we obtain perfect information is known. One criterion is based on a limiting argument where the sample size of one of the groups grows to infinity while the sample size of the other group stays fixed, whereas the second criterion is based upon conditioning on the true value of the parameters for one of the groups. In the context of testing whether 2 normal means are equal or not, we find priors where the limiting argument and conditioning give rise to equivalent Bayes decisions under perfect knowledge and cases where they give rise to different Bayes decisions. We also show that, with some prior specifications, the limiting Bayes decisions are not compatible with any prior specification for the one-sample problem where one of the distributions is known.

E0603: **Generalized variational inference**
Presenter: **Jack Jewson**, University of Warwick, United Kingdom
Bayesian inference is viewed as an optimisation problem. This is commonly associated with Variational Inference (VI) which many consider to be unprincipled and ad hoc. In fact, VI can be shown to produce the optimal posterior beliefs within the constrained family Q according to the objective function specified by Bayes rule. We use this observation to further generalise variational inference. We define optimal posterior beliefs using a triple: Q the family of admissible distributions to characterise posterior beliefs; $l(\theta,x)$ the loss function connecting observed data to the parameter of interest for the analysis; D a divergence regularising the optimal posterior beliefs towards the prior. We demonstrate how changing $l(\theta,x)$ can lead to inference that is automatically robust to outliers while changing D impacts the posterior uncertainty quantification, allowing us to improve the accuracy of estimates of marginal posterior uncertainty and produce posteriors that are less sensitive to prior specifications. Formalising variational inference in this way allows us to improve transparency and performance over the myriad of approximate inference methods which attempt to minimise different divergences between the approximate and exact posteriors.

E1329: **Computational aspects of L1-regularized g priors**
Presenter: **Christopher Hans**, The Ohio State University, United States
Many regularization priors for Bayesian regression assume the regression coefficients are a priori independent. In particular this is the case for Bayesian treatments of the lasso and the elastic net. While independence may be reasonable in some data-analytic settings, having the ability to incorporate dependence in these prior distributions would allow for greater modeling flexibility. L1-regularized g priors are one such approach to incorporating dependence and represent a special case of a general "orthant normal" prior. We investigate properties of these L1-regularized g priors and discuss efficient posterior computation. Evaluating the moment generating function of the L1-norm of a multivariate normal random vector plays a critical role in computation. We introduce bounds for this quantity and investigate several computational approaches for estimation of it. The results are applied to the problem of model comparison and variable selection in the L1-regularized g prior setting.

EO833 Room MAL G14 ADVANCES IN DISTRIBUTIONAL REGRESSION MODELS Chair: Helga Wagner

E0521: **Enhanced variable selection for boosting distributional regression**
Presenter: **Andreas Mayr**, University of Bonn, Germany
Co-authors: Annika Stroemer, Leonie Weinhold, Christian Staerk
An alternative way to fit distributional regression models is component-wise gradient boosting. Boosting leads to data-driven variable selection and works for high-dimensional data, while the resulting additive model is in the same way interpretable as if it was fitted via classical inference schemes. While being very flexible and also relatively easy to extend, in some practical applications the algorithm shows the tendency towards selecting too many variables, including false positives. This seems to take place particularly for rather low-dimensional data ($p < n$). To deal with this, we analyse different approaches to either de-select variables or stop the algorithm before it starts selecting non-informative ones. We illustrate the different approaches with a recent analysis of the health-related quality of life of patients with chronic kidney disease - using boosting to select the most informative predictors fitting a distributional beta regression model.

E0579: **Distributional trees and forests for circular data**
Presenter: **Moritz Nikolaus Lang**, University of Innsbruck, Austria
Co-authors: Lisa Schlosser, Torsten Hothorn, Georg Johann Mayr, Reto Stauffer, Achim Zeileis
Circular data can be found in a variety of applications and subject areas, e.g., hourly crime rates in the social-economics, animal movement direction in ecology, and wind direction as one of the most important weather variables in meteorology. For probabilistic modeling of circular data the von Mises distribution is widely used. While most existing approaches are built on additive regression models, we propose an adaption of regression trees to circular data by employing distributional trees. In comparison to the more commonly-used additive models, the resulting distributional trees are easy to interpret, can detect non-additive effects, and automatically select covariates and their interactions. In addition, as a natural extension, ensembles or forests of such circular trees are introduced that can further improve the forecasts by regularizing and stabilizing the model. For illustration, short-term probabilistic wind direction nowcasts at different airports are obtained in order to direct airplanes to a safe landing. The predictive skill of the novel approaches is benchmarked with an additive regression model plus a persistency and a climatology model, employing the circular continuous ranked probability score. The proposed methods for circular distributional trees and forests are available in the R package 'disttree' from R-Forge.

E0249: **Multivariate conditional transformation models**
Presenter: **Thomas Kneib**, University of Goettingen, Germany
Co-authors: Nadja Klein, Torsten Hothorn
Regression models describing the joint distribution of multivariate response variables conditional on covariate information have become an im-

portant aspect of contemporary regression analysis. However, a limitation of such models is that they often rely on rather simplistic assumptions, e.g. a constant dependency structure that is not allowed to vary with the covariates. We propose a general framework for multivariate conditional transformation models that overcomes such limitations and describes the full joint distribution in simple, interpretable terms. Among the particular merits of the framework are that it can be embedded into likelihood-based inference and allows the dependence structure to vary with the covariates. In addition, the framework scales beyond bivariate response situations, which were the main focus of earlier investigations. We illustrate the application of multivariate conditional transformation models in a trivariate analysis of childhood undernutrition and demonstrate empirically that even complex multivariate data-generating processes can be inferred from observations.

E0769: **Dynamic mixture of experts models for online predictions**
Presenter: **Parfait Munezero**, Stockholm University and Ericsson AB, Sweden
Co-authors: Mattias Villani

Mixture of experts models provide a flexible framework of modelling the density of a response variable using finite mixture models with component density functions and mixture weights depending on a set of covariates. We propose a class of dynamic mixture of experts models for online (real-time) predictions. Our model allows the component models to be any density functions not necessarily limited to the exponential family, and the parameters to vary over time. The inference is done in a Bayesian framework using sequential Monte Carlo (SMC) a.k.a Particle filter algorithms. The smoothness of parameters is controlled through a random walk prior process, which allows the parameters to fluctuate and adapt locally through time or to remain constant over time. We, therefore, propose an inference method that applies to models with either static or dynamic parameters in a unified way. We apply the model to a real dataset consisting of faults reported on a series of upgrades of a large-scale software. Further, we assess the performance of our inference method using different simulation scenarios.

EO060 Room MAL G15 TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Raffaele Argiento

E0622: **Bayesian inference in models made of modules**
Presenter: **Pierre Jacob**, Harvard University, United States
Co-authors: Chris Holmes, Christian Robert, Lawrence Murray, George Nicholson

Statisticians are faced with integrating heterogeneous data modalities relevant for an inference or decision problem. It is convenient to use a graphical model to represent the statistical dependencies, via a set of connected "modules", each relating to a specific data modality, and drawing on specific domain expertise in their development. Each module can involve parametric, semi-parametric or non-parametric components. In principle, given data, the conventional statistical update then allows for coherent uncertainty quantification and information propagation through and across the modules. However, misspecification of any module can contaminate the update of others. In various settings, particularly when certain modules are trusted more than others, practitioners have preferred to avoid learning with the full (joint) model in favor of "cut distributions". We will describe why these modular approaches might be preferable to the full model in misspecified settings, and propose criteria to choose between modular and full-model approaches. The question is intertwined with computational difficulties associated with the cut distribution.

E1236: **Importance conditional sampler for Bayesian nonparametric mixtures**
Presenter: **Riccardo Corradin**, University of Milano Bicocca, Italy
Co-authors: Bernardo Nipoti, Antonio Canale

Bayesian nonparametric mixtures are flexible models for density estimation and model based clustering, nowadays a common tool for the applied statisticians. In this family of models, Pitman-Yor mixtures show a good balance between mathematical tractability and flexibility. Inference for this class of models is mainly performed by means of MCMC methods, which can be divided into marginal and conditional methods. Marginal methods are easily interpretable, although they underestimate the uncertainty associated to posterior quantities. On the other hand, conditional methods provide an accurate estimation of posterior uncertainty. We recently introduced a sampling strategy to estimate Pitman-Yor mixtures, named importance conditional sampler (ICS). Our proposal has proved to be highly efficient and robust to the specification of the parameters characterising the distribution of the underlying process. The ICS provides an accurate estimation of posterior uncertainty, and, like marginal methods, it is described by a simple and interpretable predictive structure. Motivated by an astronomical application, we extended the ICS approach to Griffiths-Milne dependent mixtures, a family of models for partially exchangeable data.

E1098: **Nonparametric Bayesian inference and goodness of fit testing for stochastic differential equations**
Presenter: **Ioanna Manolopoulou**, University College London, United Kingdom
Co-authors: Yvo Pokern, Tjun Yee Hoh

Methodology is developed for non-parametric Bayesian inference for 2-dimensional diffusion processes, motivated by the study of trajectories in animal movement. We employ an interpretable conjugate Gaussian measure prior whose precision operator is chosen to be a high order differential operator, and construct efficient pseudo-spectral samplers for Markov chain Monte Carlo posterior sampling for the drift, diffusivity and diffusion bridges. Evaluation of model fit for diffusion processes is currently lacking, with many applications of diffusions to real data in the literature suffering obviously poor model fit. We extend an existing transition density-based test to the case of non-parametric drift. We study the finite-sample behaviour of the test statistic and compute Bayesian discrepancy p-values. We illustrate our methods on a dataset following the movement of a Capuchin monkey and describe how outlier removal and systematic sub-sampling of the data can be beneficial to model fit.

E0587: **Bayesian capture-recapture models with temporary emigration and heterogeneity using mixtures of changepoint processes**
Presenter: **Eleni Matechou**, University of Kent, United Kingdom
Co-authors: Raffaele Argiento

In populations where individuals are detected with probability lower than one, capture-recapture methodology is often employed for estimating population size and, in open populations, arrival and departure times. However, all the existing capture-recapture models either assume that emigration is permanent and each individual performs a single visit to the site of interest, or require the use of specialised sampling schemes that assume population closure for certain times of the sampling period. We propose a novel approach for modelling capture-recapture data on open populations that exhibit temporary emigration, whilst also accounting for individual heterogeneity. Our modelling approach combines changepoint processes, fitted using an adaptive approach, for inferring the number and timing of individual visits with Bayesian mixture modelling, fitted using a nonparametric approach, for identifying clusters of individuals with similar visit patterns and capture probabilities. The proposed method is extremely flexible as it can be applied to any capture-recapture data set and is not reliant upon specialised sampling schemes. We demonstrate our model and algorithm using a motivating data set on anglers fishing in the Gaula river in Norway and the results provide us with the first ever estimate of the size of the population of anglers fishing during the season as well as new insights on the visit patterns of anglers and their probabilities of catching salmon whilst at the river.

EO438 Room MAL G16 EAS SESSION: EMERGING APPLICATIONS WITH COPULAS Chair: Richard Everitt

E0330: **Time-varying copula models for longitudinal data**
Presenter: **Esra Kurum**, University of California, Riverside, United States

A copula based joint modeling framework for mixed longitudinal responses is proposed. The approach permits all model parameters to vary with time, and thus will enable researchers to reveal dynamic response-predictor relationships and response-response associations. We call the new class of models time-cop because we model dependence using a time varying copula. We develop a one step estimation procedure for the time-cop

parameter vector, and also describe how to estimate standard errors. We investigate the finite sample performance of our procedure via simulation studies, one of which shows that our procedure performs well under ignorable missingness. We also illustrate the applicability of our approach by analyzing binary and continuous responses from the Women's Interagency HIV Study.

E0395: **Bayesian nonparametric approach for vine copula modelling: An application to preterm birth data in repeated pregnancies**
Presenter: **Rosario Barone**, Sapienza University of Rome, Italy
Co-authors: Luciana Dalla Valle

Preterm births represent a serious medical issue since they can affect the health of the mother and the fetus. Since women tend to be affected by adverse outcomes in repeated pregnancies, we focus on the study of the dependence between preterm births in repeated pregnancies using a vine copula approach. Our dataset includes 164 women with gestational ages lower than 37 weeks, each of whom has had at least three pregnancies. Since we are working with truncated data, distributional assumptions on the margins might be restrictive: we model the gestational age of each pregnancy as a GAMLSS (more specifically as a truncated Weibull), choosing as covariates information about smoking, history of preterm birth, parity and type of preterm birth. Then, we follow a Bayesian nonparametric method to estimate the pair copulas in the vine, extending the approach presented by Wu et. al (2015), which used an infinite mixture of Gaussian copula densities to define a nonparametric copula for modelling any dependence structure between the marginals, to the vine copula setting. Therefore, we assume a Dirichlet process prior on each pair copula. Our approach has two main advantages compared to the traditional methods: on the one hand it is extremely flexible, due to the vine structure, and on the other hand it overcomes the need of specify the families of each pair copula.

E0424: **A catastrophe model for insurance losses due to freeze events using vine copulas**
Presenter: **Symeon Koumoutsaris**, Guy Carpenter, United Kingdom

The probability of extreme cold weather events in the United Kingdom is assessed as part of a probabilistic catastrophe model for insurance losses. Reanalysis data have been employed to construct the hazard part of the model, which is based on the Air Freezing Index, an index which accounts for both the magnitude and the duration of air temperature below freezing. Extreme value analysis is used to obtain return level predictions on a 1x1 degree grid. More importantly, the spatial dependence of the hazard between the grid cells has been assessed through a novel approach which takes advantage of the vine copula methodology. This approach allows the modelling of concurrent high AFI values across the country, which is necessary in order to assess the extreme behaviour of freeze events. Recognizing the non-stationary nature of climate extremes, the model also incorporates the effects of the North Atlantic Oscillation and climate change effects in the occurrence of cold spells. However, considerable uncertainty exists in these results, owing mainly to the short record length and the large interannual variability of the AFI.

E0502: **Objective interpretation of the penalised complexity prior in copula models**
Presenter: **Diego Battagliese**, University of Rome La Sapienza, Italy
Co-authors: Clara Grazian, Brunero Liseo, Cristiano Villa

In many statistical models it is natural to have a nested structure. Consider a model of a given complexity, one way to obtain a more flexible model is to include an extra component so that the simpler model would be nested in the more complex one. One may think, for instance, of a situation where one wants to model the joint distribution of several random variables through a copula function. In the case of dependence among variables, the joint density can be expressed as the product of the marginal distributions times a copula function, on the contrary, the joint density boils down to the only product of the marginals when the variables are independent. For the Gaussian copula model we derive a Penalised Complexity prior. We also show that for any copula function and for any dimension the elicitation of the PC prior is invariant with respect to the marginals and their parameters, indeed only the copula structure matters. PC priors are constructed by means of a user-defined scaling, but in order to be objective we assign an intrinsic prior to the scale parameter of the PC prior. Then, objectivity is attained by maximizing the variance of the PC prior where an intrinsic prior is put on the scale parameter. The PC prior can be used in Bayesian hypothesis testing both in the calculation of the Bayes factor and in the calibration of objective prior distributions on the models. To check the goodness of the prior, a comparison with other existing priors is carried out.

EO436 Room CLO 101 FUNTIONAL DATA ANALYSIS AND DEPENDENT SEQUENCES Chair: Anne Francoise Yao

E1036: **A study of the manifold hypothesis for functional data by using spectral clustering**
Presenter: **Julien Ah-Pine**, University of Lyon, France
Co-authors: Anne Francoise Yao

Most of functional data clustering methods assume that the observations belong to linear subspaces. This hypothesis may not be verified in practice. To investigate this point we use spectral clustering on functional data. This clustering method uses the eigen-decomposition of the (discrete) Laplacian of the affinity graph of the observations as a Euclidean embedding of the proximity relationships. In this framework, several affinity measures and neighborhood selection procedures can be used in order to approach the non-linear manifold underlying the data. Our experimental results include several real-world clustering tasks and support the manifold hypothesis for functional data.

E1050: **Bayesian inference for shape and high-dimensional inputs with a Gaussian process prior**
Presenter: **Chafik Samir**, UCA-LMBP/CNRS, France
Co-authors: Anis Fradi

A Bayesian point of view is adopted for studying shapes and high-dimensional data. Since computing the exact marginal likelihood for a Bayesian model remains difficult if not impossible in high-dimensional and manifolds data, we introduce two types of methods to improve the efficiency and the scalability of Gaussian processes. We to handle multimodal posteriors and the prediction quality, jointly. We furthermore show some asymptotic properties such as non-unbiasedness, prediction sufficiency, etc. The proposed approaches are shown to provide computational advantages with respect to some existing methods that rely on modified covariance function. Various tests on real and simulated data will be discussed as well as the efficiency of Bayesian and non-Bayesian predictors.

E1060: **A data-driven smooth test of comparison for dependent sequences**
Presenter: **Laurence Reboul**, Aix-Marseille Universite, France
Co-authors: Denys Pommeret, Anne Francoise Yao

A smooth test of comparison for the marginal distributions of strictly stationary dependent bivariate sequences is proposed. We first state a general test procedure and several cases of dependence are then investigated, including short and long memory cases. The test is applied to both simulated data and real datasets and achieves good performances.

E1276: **Kernel regression estimation for functional predictor with values in a Riemannian submanifold**
Presenter: **Anne Francoise Yao**, Universite Clermont Auvergne/LMBP, France
Co-authors: Chafik Samir, Papa Mbaye

The problem of kernel estimation of the regression function of a real valued random variable Y given a functional one X will be considered. We focus on the behavior of the Nadaraya-Watson in the situation where X is with values in a finite dimensional Riemannian submanifold of a Hilbert space. We illustrate our purpose through some real data applications.

EO536 Room CLO 102 ADVANCES IN COMPLEX DATA MODELING Chair: Maria Francesca Marino

E0434: **Parametric modelling of quantile functions**
Presenter: **Paolo Frumento**, Karolinska Institute, Sweden
Co-authors: Matteo Bottai
Quantiles can be used to describe complexity and diversity, and to capture features beyond location and scale parameters. In traditional quantile regression, no parametric structure is imposed and different quantiles are estimated one at a time. While this is generally seen as an advantage, it also presents a number of important drawbacks: (i) it is statistically inefficient; (ii) it increases the chance of quantile crossing; (iii) it generates a large amount of information, making it difficult to summarise and interpret the results; (iv) it does not allow using identifying assumptions; and (v) it makes it difficult to apply quantile regression to censored, truncated, or longitudinal data. Describing the quantile function by a parsimonious parametric model represents an obvious, yet unexplored alternative. A new, broad family of models and estimators has been designed and implemented. An R package 'qrcm' provides all the necessary functions for inference, plotting, prediction, and goodness-of-fit assessment.

E1058: **Recent advances in iterative imputation models**
Presenter: **Gerko Vink**, Utrecht University, Netherlands
Co-authors: Stef van Buuren
People are nowadays aware that not treating missing data problems may be the least desirable approach to drawing inference from incomplete data. Many routines to handle missing values are therefore included in standard statistical software. Most of the incomplete data problems are unfortunately not standard and require approaches that go beyond the standard solutions: some problems would benefit from a multivariate imputation approach, while univariate imputation would be more flexible for other problems. Recent advances in iterative imputation are detailed that allow for closer modeling of the imputation problem at hand. The focus lies on hybrid imputation, where univariate and multivariate imputation methods are combined through the specification of imputation blocks. We will explore the theory and application of some of these newer imputation models.

E1273: **Algorithms and diagnostics for the analysis of ranking data with the extended Plackett-Luce model**
Presenter: **Cristina Mollica**, Sapienza Universita di Roma, Italy
Co-authors: Luca Tardella
The Plackett-Luce distribution (PL) is one of the most successful parametric options within the class of multistage ranking models to learn preferences on a given set of items from a sample of ranked sequences. It postulates that the ranking process is carried out by sequentially assigning the positions according to the forward order, that is, from the top (most-liked) to the bottom (least-liked) alternative. This assumption has been recently relaxed with the Extended Plackett-Luce model (EPL) with the introduction of the discrete reference order parameter, describing the rank attribution path. By starting from the formal proof of a special property of the EPL, related to the inverse ordering of the item probabilities at the first and last stage of the ranking process, we derive a diagnostic tool for the EPL distribution whose inferential utility is motivated from a double perspective. First, the novel statistic is proposed to test the appropriateness of the EPL assumption. In this regard, the new diagnostic contributes to fill the deficiency of goodness-of-fit methods for the family of multistage models. Besides model adequacy evaluation, we also show how the statistic can be exploited to construct a heuristic method that surrogates the likelihood approach for inferring the underlying reference order parameter. The usefulness of the proposals is illustrated with a simulation study and applications to real ranking data.

E1297: **Adjusting for a pre-test in school value-added models: A comparison between gain score and conditional approaches**
Presenter: **Carla Rampichini**, University of Florence, Italy
Co-authors: Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto
The issue of prior achievement adjustment in value added models is considered. Based on data of the National Institute for the Evaluation of the Education System in Italy (INVALSI), our focus is on the effect of the lower secondary school type (public versus private) on test scores at the 8th grade (post-test), accounting for students test scores at the 5th grade (pre-test). Such effect can be estimated by either adjusting for the pre-test score (i.e. conditioning) or by using the difference between post-test and pre-test scores (gain score) as response variable. We discuss the assumptions underlying the gain score approach as compared to the conditional approach, and the consequences of their violations. We evaluate the two approaches by means of an application using INVALSI data and by a simulation study that explicitly takes into account the multilevel structure of the data. The results show that the performance of the two approaches, in terms of bias and efficiency, depends on several factors, such as pre-test reliability and validity of the common trend assumption.

EO304 Room Jessel OUTLIERS AND STRUCTURAL BREAKS Chair: Alexander Duerre

E1436: **Nonparametric test for heteroscedasticity in time series**
Presenter: **Herold Dehling**, Ruhr-University Bochum, Germany
Co-authors: Roland Fried, Sara Kristin Schmidt, Max Wornowizki
A new test is presented for heteroscedasticity of a time series that is based on recent ideas. The test statistic is given by Gini's mean difference of logarithmic local sample variances. We investigate the asymptotic distribution of our test statistic under the null hypothesis of constant variance, and establish consistency against a large class of alternatives. Our results hold under mild conditions concerning the dependence structure of the underlying time series.

E0924: **Change point detection by distance covariance function**
Presenter: **Konstantinos Fokianos**, Lancaster University, United Kingdom
The focus is on the problem of non-parametric change-point detection for multivariate time series data in the sense of discovering changes in the marginal distribution of the process. We employ a methodology that is based on properties of the distance covariance function. The main ideal is to investigate the behavior of a cumulative sum type of process in terms of the characteristic function. In fact, the test statistic we consider has a closed form expression and can be easily implemented to any given data stream. Some preliminary results show the validity of this approach.

E1480: **On the finite sample behaviour of the cusum test**
Presenter: **Alexander Duerre**, TU Dortmund, Germany
The cusum test is a very popular tool in nonparametric change point analysis. It is applied to decide whether an observed time series remains stationary or exhibits a structural change. Under the null hypothesis of no change, the cusum trajectory converges under very general conditions against a Brownian bridge. If there is one level shift, the trajectory gets more triangular shaped and attains larger absolute values, which is the reason one usually looks at the maximal absolute value of the cusum trajectory. This value converges to the Kolmogorov-Smirnov distribution. However, the test statistic is seriously biased which leads to conservative tests under the null hypothesis and a loss of power under the alternative. We show that this distortion, under independent normally distributed random variables, is of order $n^{(-1/2)}$ where n is the sample size. There is a straightforward correction, and we show that this correction is also reasonable in case of other symmetric distributions. Serial dependence results in even more conservative behaviour. We investigate the error under autoregressive processes.

E1885: **Estimation of the long run variance in case of multiple level shifts**
Presenter: **Sheila Goerz**, TU Dortmund, Germany

Co-authors: Alexander Duerre

The cusum test is one of the most popular tools in change-point detection. Under short range dependence and fairly mild technical conditions cusum type tests depend only on one nuisance parameter, often called long run variance, but apart from that, they are distribution free. We propose a new kernel-type estimator for the long run variance. It is based on the sum of weighted autocovariances. The classical kernel or Bartlett estimator uses a weighted sum of autocovariances giving larger time lags smaller weights. We substitute autocovariance estimations which are especially suitable under change points for the common used empirical autocovariances. To be robust against level shifts we use a rolling window to estimate the location instead of the global mean. Depending on the window size the resulting estimator is highly biased. Therefore a correction is applied. In a simulation study we compare our estimator with commonly used ones.

EO248 Room MAL 152 RECENT DEVELOPMENTS IN STATISTICAL MULTISCALE METHODS Chair: Michael Vogt

E1051: **Quadratic forms of lifting transforms: Spectral analysis and stationarity tests for time series with missing values**

Presenter: **Guy Nason**, Imperial College, London, United Kingdom

Co-authors: Marina Knight, Matthew Nunes, Kathryn Leeming

Quadratic forms are ubiquitous and intensively studied in statistics, often in time series analysis, including those formed out of wavelet coefficients. Most wavelet transform methods in statistics assume regularly-spaced and complete data, which does not always occur in real problems where observations are sometimes missing, resulting in a non-regular design. We use second-generation wavelets (lifting) which are explicitly designed to handle non-regular situations. We introduce a new estimator of the second-generation wavelet spectrum and show that it is consistent in the case of an underlying locally stationary wavelet process where the observations are subject to a random drop-out model. Our new estimator is then used to construct a new lifting-based stationarity test with significance assessed by the bootstrap. The simulation study shows excellent results, not only on time series with missing observations, but in the complete case too.

E0857: **Tests for qualitative features in the random coefficients model**

Presenter: **Fabian Dunker**, University of Canterbury, New Zealand

Co-authors: Konstantin Eckle, Katharina Proksch, Johannes Schmidt-Hieber

The linear random coefficient model $Y_i = \beta_{i,1}X_{i,1} + \beta_{i,2}X_{i,2} + \ldots + \beta_{i,d}X_{i,d}$ is an effective way to model unobserved heterogeneity. Here $(\mathbf{X}_i, Y_i)$, $i = 1,\ldots,n$, are i.i.d. observations with $\mathbf{X}_i = (X_{i,1},\ldots,X_{i,d})$ being a d-dimensional vector of regressors and Y_i a univariate responds. The random coefficients $\beta_i = (\beta_{i,1},\ldots,\beta_{i,d})$, $i = 1,\ldots,n$ are unobserved i.i.d. realizations of d-variate random vector with unknown density f_β independent of $\mathbf{X}_i$. We propose and analyze a nonparametric multiscale test for slopes and modes of the random coefficient density f_β. The test uses the connection between the model and the d-dimensional Radon transform and is based on Gaussian approximation of empirical processes.

E0799: **Statistical multiscale analysis for molecules**

Presenter: **Katharina Proksch**, University of Twente, Netherlands

Co-authors: Frank Werner, Jan Keller

In recent years, many new super-resolution microscopy techniques such as STED or RESOLFT have been developed such that inference on the level of single molecules is now a reasonable goal in fluorescence microscopy. We discuss options to infer on the number and locations of molecules in a given sample based on a hybrid algorithm, which offers both the segmentation of an image based on multiscale methods as well as estimates of the local numbers of molecules, while it preserves uniform confidence about all statements. The proposed method allows for the construction of a novel, automatized statistical analysis tool for scanning microscopy via a molecular map, that is, a graphical presentation of locations as well as local numbers of molecules and corresponding uniform confidence statements. It is based on a sound statistical model, which connects both the local brightness and molecule distributions with the fact that a single molecule can emit only one photon at a time (antibunching). More precisely, our method is built on rigorous statistical convolution modeling of higher order photon coincidences and an approach on hot spot detection in heterogeneous data via multiscale scan statistics. We demonstrate the functionality of the molecular map by means of data examples from STED fluorescence microscopy.

E0882: **Multiscale inference for nonparametric time trends**

Presenter: **Michael Vogt**, University of Bonn, Germany

Co-authors: Marina Khismatullina

New multiscale methods are developed to test qualitative hypotheses about the regression function m in a nonparametric regression model with fixed design points and time series errors. In time series applications, m represents a nonparametric time trend. Practitioners are often interested in whether the trend m has certain shape properties. For example, they would like to know whether m is increasing/decreasing in certain time intervals. The multiscale methods allow us to test for such shape properties of the trend m. In order to perform the methods, we require an estimator of the long-run variance of the error process. We propose a new difference-based estimator of the long-run error variance for the case that the error terms have an autoregressive structure. The usefulness of our methods is illustrated by an empirical application to climate data.

EO058 Room MAL 153 RECENT ADVANCES IN QUANTILE REGRESSION Chair: Anneleen Verhasselt

E0426: **A new quantile regression with application to the analysis of bounded variables**

Presenter: **Keming Yu**, Brunel University, United Kingdom

A new quantile regression model will be presented for the analysis of bounded dependent variables such as well-being scores and the Gini coefficients.

E0509: **Non-crossing p-splines quantile regression on time varying coefficient model of crude palm oil production in Indonesia**

Presenter: **Yudhie Andriyana**, Universitas Padjadjaran, Indonesia

The use of nonparametric approaches in a quantile regression technique is important, because some parts of quantile levels are very hard to specify parametrically. Unfortunately, similar to the parametric approach, the estimated regression quantile curves using nonparametric techniques often cross each other, which can be very annoying for interpretations and further analysis. We are concerned with a non-crossing p-splines quantile objective function on time varying-coefficient modelling. The performances of the proposed techniques is investigated via some simulation studies and applied to crude palm oil (CPO) data as one of the main export commodities in Indonesia. By means of non-crossing p-splines quantile objective function, we model a growth pattern of CPO involving some covariates, such as, rainfall, temperature and humidity.

E0537: **Expectation propagation for generalised quantile regression models**

Presenter: **Mauro Bernardi**, University of Padova, Italy

L_α–quantile regression models generalise quantiles ($\alpha = 1$) and expectiles ($\alpha = 2$) regression to account for the whole conditional distribution of the response variable. We introduce the L_α–quantile regression model and we present a new Bayesian estimation framework where regression parameters are learned by minimising the expected tilted check function. An approximated model evidence is obtained by employing the Expectation Propagation (EP) algorithm. The analytically intractable integration required by the parameters learning problem is solved by minimising the Kullback–Leibler divergence between the unnormalised posterior and a suitable approximating distribution usually belonging to the exponential family. We also provide some theoretical results concerning the consistency of the posterior distribution of the regression parameters under general priors. Moreover, the model selection problem is approached through an approximated Stochastic Search Variable Selection (SSVS–EP) algorithm

based on the spike–and–slab prior. The effectiveness of the proposed model and parameter learning method are assessed on synthetic and real datasets.

E0542: **Semiparametric quantile regression using quantile-based asymmetric family of densities**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium
Co-authors: Irene Gijbels, Md Rezaul Karim

Quantile regression is an important tool in data analysis. Linear regression, or more generally, parametric quantile regression often imposes too restrictive assumptions. Nonparametric regression avoids making distributional assumptions, but might have the disadvantage of not exploiting distributional modeling elements that might be brought in. We discuss a semiparametric approach towards estimating conditional quantile curves. It is based on a recently studied large family of asymmetric densities for which the location parameter is a quantile (and not a mean). Passing to conditional densities and exploiting local likelihood techniques in a multiparameter functional setting then leads to a semiparametric estimation procedure. Due to the appealing semiparametric framework, we can discuss in detail the bandwidth selection issue, and provide several practical bandwidth selectors. The practical use of the semiparametric method is illustrated in the analysis of maximum winds speeds of hurricanes in the North Atlantic region.

EO550 Room MAL 254 SPATIAL INFERENCE Chair: Marco Meyer

E0898: **Finite predictor coefficients and the inverse Yule-Walker matrix: On the extension of Akaike's identity to random fields**
Presenter: **Carsten Jentsch**, TU Dortmund University, Germany
Co-authors: Marco Meyer

For univariate stationary and centered time series $(X_t)_{t\in\mathbb{Z}}$, a useful identity links the inverse of the Yule-Walker matrix $\Gamma(p) = E(XX')$, where $X = (X_{t-1},\ldots,X_{t-p})'$, to the corresponding finite predictor coefficients. This factorization of $\Gamma(p)^{-1}$ is employed in different areas of statistics, and it is particularly crucial to derive asymptotic theory for autoregressive spectral density estimators. We investigate the validity of a natural extension of this factorization to univariate stationary random fields on a lattice $(X_t)_{t\in\mathbb{Z}^d}$. We prove the surprising result that such a factorization holds true if and only if a certain Toeplitz condition on the autocovariance function $\gamma(h) = E(X_{t+h}X_t)$ and the shape of the fitted autoregressive models is fulfilled. We show that this condition turns out to be very restrictive. In fact, it implies that an analogue of Akaike's identity in general does not hold for many commonly used spatial autoregressive models such as half-plane or quarter-plane models. Instead, we establish an unexpected link between the entries of the inverse Yule-Walker matrix and prediction theory. We illustrate the theoretical findings for different kinds of autoregressive fits.

E0329: **A general frequency domain method for assessing spatial covariance structures**
Presenter: **Soutir Bandyopadhyay**, Colorado School of Mines, United States
Co-authors: Soumendra Lahiri, Daniel Nordman

In examining dependence in spatial data, it can be helpful to assess hypotheses about spatial covariance that may not be fully model-based or parametric. That is, one may wish to test for general features regarding spatial covariance without presupposing any particular, or potentially restrictive, assumptions about the joint data distribution. Current methods for testing spatial covariance are often intended for specialized inference scenarios, usually with spatial lattice data. We propose instead a general method for estimation and testing of spatial covariance structure, which is valid for a variety of inference problems (including nonparametric hypotheses) and applies to a large class of spatial sampling designs with irregular data locations. The proposed method has the advantage of providing valid inference in the frequency domain without estimation of such standard errors, which are often intractable, and without particular distributional assumptions about the data (e.g., Gaussianity). To illustrate that, we develop the method for formally testing isotropy and separability in spatial covariance and consider confidence regions for spatial parameters in variogram model fitting. A broad result is also presented to validate the method for further general tests of spatial covariance structure. The approach uses spatial test statistics, based on an extended version of empirical likelihood, having simple chi-square limits for calibrating tests.

E0566: **A general central limit theorem and subsampling variance estimator for alpha-mixing multivariate point processes**
Presenter: **Christophe Biscio**, Aalborg University, Denmark

Central limit theorems for multivariate summary statistics of alpha-mixing spatial point processes have usually been established using either the so-called Bernstein's blocking technique or an approach based on Bolthausen's results. It is characteristic that essentially the same theorems have been (re)-invented again and again for different specific settings and statistic considered. Moreover, although there exists estimates in some particular cases, the asymptotic variance is usually unknown or difficult to compute. We present a unified framework based on Bolthausen's work to state, once and for all, a general central limit theorem for alpha-mixing multivariate point process that applies in a general non-stationary setting and is also applicable to non-parametric kernel estimators depending on a bandwidth converging to zero. In particular, we argue why this approach is more suitable than the one using Bernstein's blocking technique. We believe this can save a lot of work and tedious repetitions in future applications of alpha-mixing point processes.

E0794: **Spatial GARCH-type models: A unified approach**
Presenter: **Philipp Otto**, Leibniz University Hannover, Germany
Co-authors: Wolfgang Schmid

In time series analysis and, particularly, in finance, (generalized) autoregressive conditional heteroscedasticity models are widely applied statistical tools for modelling volatility clusters, i.e., periods of increased or decreased risks. In contrast, spatial dependence in the conditional second moments of spatial and spatiotemporal processes has been seen rather uncritical up to now. There are only a few models, which have been proposed for modelling local clusters of increased risks. We introduce a unified spatial and spatiotemporal GARCH-type model, which covers all previously proposed spatial ARCH models but also introduces novel spatial GARCH as well as E-GARCH processes. For this common modelling framework, maximum-likelihood estimators are derived. In addition to the theoretical contributions, we suggest a model selection strategy verified by a series of Monte-Carlo simulation studies. Eventually, the use of the unified model is demonstrated by an empirical example. In particular, we focus on real-estate prices from 1995 to 2014 in all Berlin ZIP code areas. For these data, a spatial autoregressive model has been applied, which shows locally varying model uncertainties captured by the spatial GARCH-type models.

EO112 Room Senate BAYESIAN INFERENCE FOR EXTREME VALUES Chair: Miguel de Carvalho

E0248: **Modeling exceedances in extreme value theory: Foundations, regression, time series and multivariate settings**
Presenter: **Dani Gamerman**, Universidade Federal do Rio de Janeiro, Brazil

Extreme value theory (EVT) is the branch of Statistics concerned with extremes or tails of a distribution. It has a long list of areas of application, including Finance and Environmental Sciences. One of the main concerns of EVT is to model exceedances, or values beyond a sufficiently high quantile. Nice theoretical results suggest the way forward to approximate exceedance behaviour, but do not define how extreme one needs to be for the approximation to work well. Ad-hoc procedures are commonly used to address this issue but they suffer from the pitfalls inherent to such procedures and do not take into account the uncertainty associated. Thus, resulting inference is likely to be biased and/or to underestimate uncertainty. We propose a procedure that avoids such pitfalls by letting the data drive the decision of when the approximation can be safely applied,

while accounting for the uncertainty of this choice. The procedures are extended to: 1) accommodate for the inclusion of external sources of information; 2) the time series context to incorporate temporal dependence; 3) identify the extremal behavior, and;4) handle multivariate contexts.

E0307: **Flexible modeling for spatial extremes with application to environmental studies**
Presenter: **Yuan Tian**, North Carolina State University, United States
Co-authors: Brian Reich

Extreme events, such as unusually high temperature, major precipitation and life-threatening hurricanes, occur with an extremely small probability, but may have catastrophic consequences. It is therefore of great significance to make inferences and predictions about these rare events. Extreme value analysis plays a major role in modeling these rare events. One common approach is the max-stable process. A Bayesian hierarchical structure has been previously proposed that can be easily implemented and efficiently computed via MCMC algorithm to present the max-stable process. We extend the current RS model using nonparametric Bayesian modeling that relaxes the assumptions. Thus, we provide more flexibility in expanding the class of marginal distributions. In addition, we present a hybrid model that combines the strength of the original RS model and the nonparametric model. Tail behavior of the novel model is studied. The utility of the proposed model is evaluated in Monte Carlo simulation studies and is applied to precipitation data for the US.

E0404: **Data imputation of large observations via Bayesian inference for multivariate extremes**
Presenter: **Isadora Antoniano-Villalobos**, Ca' Foscari University of Venice, Italy
Co-authors: Simone Padoan

Missing data is a known issue in statistics. In many environmental applications, the greatest interest is placed on large observations, e.g. of pollution levels, wind speed, precipitation or temperature, to name a few. In such contexts, usual data imputation methods may fail to reproduce the heavy tail behaviour of the quantities involved. Recent literature has proposed the use of multivariate extreme value theory to predict an unobserved component of a random vector given large observed values of the rest. This is achieved through the estimation of the angular measure controlling the dependence structure in the tail of the distribution. The idea can be used for effective data imputation at adequately large levels, provided that the model used for the angular measure is flexible enough to capture complex dependence structures. A Bayesian nonparametric model based on constrained Bernstein polynomials ensures such flexibility, while allowing for tractable inference. An additional advantage of this approach is the natural way in which uncertainty about the estimation is incorporated into the imputed values through the Bayesian paradigm.

E0482: **Modeling time-changing joint extremes via Bernstein polynomials**
Presenter: **Timothy Hanson**, Medtronic, United States
Co-authors: Miguel de Carvalho

A Bayesian model for monitoring the dynamics ruling the dependence between extreme values over time is developed and validated. The model is based on a mean-constrained Bernstein polynomial that can be used to induce a prior on the space of all time-changing extreme value copulas. Practical notes on obtaining and interpreting inference are presented. We examine the performance of the model via a simulation study, and illustrate the methods with a real data case study.

EO588 Room CLO 203 TWO PHASE DESIGNS FOR CORRELATED DATA Chair: Babette Brumback

E1346: **Outcome-dependent sampling in cluster-correlated settings with application to hospital profiling**
Presenter: **Glen McGee**, Harvard University, United States
Co-authors: Jonathan Schildcrout, Sharon-Lise Normand, Sebastien Haneuse

Hospital readmission is a key marker of quality of health care used by the Centers for Medicare and Medicaid Services to determine hospital reimbursement rates. Analyses of readmission are based on a logistic-normal generalized linear mixed model (GLMM) that permits estimation of hospital-specific measures while adjusting for case-mix differences. Moves to address healthcare disparities call for expanding adjustment to include socioeconomic measures while minimizing burden to hospitals associated with data collection. We propose the detailed socioeconomic data to be collected on a sub-sample of patients via an outcome-dependent sampling scheme, specifically the cluster-stratified case-control (CSCC) design. Estimation and inference, for both fixed and random effects components, is performed via pseudo-maximum likelihood wherein inverse-probability weights are incorporated into the integrated likelihood to account for the design. In simulations, CSCC sampling proves to be an efficient design whenever interest lies in fixed or random effects of a GLMM and covariates are unobserved or expensive to collect. Methods are illustrated via a motivating analysis of Medicare beneficiaries hospitalized for congestive heart failure at one of 3,116 hospitals. Results show that the proposed framework provides a means of mitigating disparities in terms of which hospitals are indicated as being poor performers, relative to a naive analysis that fails to adjust for missing case-mix variables.

E0654: **Optimal sample allocation for two-phase designs in cluster correlated data settings**
Presenter: **Claudia Rivera-Rodriguez**, The University of Auckland, New Zealand
Co-authors: Sebastien Haneuse

Large amount of the research in survey sampling has been directed towards improving estimates using information available for the entire population. The efficiency of such methods depends on several factors such as the information available for the entire population, the sampling strategy, the sample size, etc. There is no an absolutely optimal design, but under certain principles and restrictions, a well designed sampling strategy can be implemented. In two-phase designs, efficiency is gained with stratification by auxiliary information known early in the design. There is a number of reasons why investigators may want to stratify: it offers gains in efficiency when the target variable behaves differently between strata and estimates can be obtained for each strata. A further way to gain efficiency is by optimally allocating the resources. For example, conditional on a given sample size or a given precision, what is the optimal allocation of sample sizes? Large amount of the research on optimal allocation has been directed towards estimation of totals or functions of totals. However, in many instances inference is concerned with regression parameters from data that arises from a correlated setting. We examine the impact of ignoring the correlation in when allocating thee resources. Using theory from sampling survey, we extend and propose different allocation methods that allow for correlated data.

E0650: **Cluster-based outcome-dependent sampling in resource-limited settings: Inference in small-samples**
Presenter: **Sebastien Haneuse**, Harvard TH Chan School of Public Health, United States
Co-authors: Sara Sauer, Bethany Hedt-Gauthier, Claudia Rivera-Rodriguez

Outcome-dependent sampling is an indispensable tool for carrying out cost-efficient research in resource-limited settings. One such sampling scheme is a cluster-based design where clusters of individuals (e.g. clinics) are selected, in part at least, on the basis of the outcome rate of the individuals. For a given dataset collected via a cluster-based outcome-dependent sampling scheme, it has been proposed to perform estimation for a marginal model using inverse-probability-weighted generalized estimating equations, where the cluster-specific weights are the inverse probability of the clinic's inclusion in the sample. We provide a detailed treatment of the asymptotic properties of this estimator, together with an explicit expression for the asymptotic variance and a corresponding estimator. Furthermore, motivated by a study we conducted in Rwanda, we provide expressions for small-sample bias corrections to the both the point estimates and the standard error estimates. Through simulation, we show that applying these corrections when the number of clusters is small generally reduces the bias in the point estimates, and results in closer to nominal coverage. The proposed methods are illustrated using data from 18 health centers in Rwanda, collected via a cluster-based outcome-dependent sampling scheme, with the goal of examining risk factors for low birth weight.

E0720: **Semiparametric generalized linear models: Application to biased samples**
Presenter: **Paul Rathouz**, University of Texas at Austin, United States

A novel class of generalized linear models indexed by a linear predictor and a link function for the mean of $(Y|X)$ has been previously proposed. In this class, the distribution of $(Y|X)$ is left unspecified and estimated from the data via exponential tilting of a reference distribution, yielding a response model that is a member of the natural exponential family. We focus on how, with very easy-to-implement modifications, the model can accommodate biased samples arising from two-phase extensions of case-control designs to count or continuous response distributions, wherein inferences about the mean are of interest.

EO368 Room CLO 204 ADVANCE IN STATISTICAL METHODS FOR LARGE AND COMPLEX DATA Chair: Dehan Kong

E0908: **Estimation of expected Euler characteristic curves of nonstationary Gaussian random fields**
Presenter: **Fabian Telschow**, University of California San Diego, United States
Co-authors: Armin Schwartzman, Dan Cheng, Pratyush Pranav

The expected Euler characteristic (EEC) curve of excursion sets of a Gaussian random field is used to approximate the distribution of its supremum for high thresholds. Viewed as a function of the excursion threshold it is expressed by the Gaussian kinematic formula (GKF) as a linear function of the Lipschitz-Killing curvatures (LKCs) of the field, which solely depend on the domain and covariance function of the field. So far its use for non-stationary Gaussian fields over non-trivial domains, has been limited because in this case the LKCs are difficult to estimate. Consistent estimators of the LKCs are proposed as linear projections of "pinned" observed Euler characteristic curves and a linear parametric estimator of the EEC curve is obtained, which is more efficient than its nonparametric counterpart for repeated observations. A multiplier bootstrap modification reduces the variance of the estimator, and allows estimation of LKCs and EEC of the limiting field of non-Gaussian fields satisfying a functional CLT. The proposed methods are evaluated using simulations and applications are presented, e.g., 3D fMRI brain activation.

E1075: **A Potts-mixture spatio-temporal joint model for combined MEG and EEG data**
Presenter: **Yin Song**, University of Victoria, Canada
Co-authors: Farouk Nathoo, Arif Babul

A new methodology is developed for determining the location and dynamics of brain activity from combined magnetoencephalography (MEG) and electroencephalography (EEG) data. The resulting inverse problem is ill-posed and is one of the most difficult problems in neuroimaging data analysis. We propose a solution that combines the data from three different modalities, MRI, MEG, and EEG, together. We propose a new Bayesian spatial finite mixture model that builds on a previous mesostate-space model. Our new model incorporates two major extensions: (i) We combine EEG and MEG data together and formulate a joint model for dealing with the two modalities simultaneously; (ii) we incorporate the Potts model to represent the spatial dependence in an allocation process that partitions the cortical surface into a small number of latent states termed mesostates. We formulate the new spatiotemporal model and derive an efficient procedure for simultaneous point estimation and model selection based on the iterated conditional modes algorithm combined with local polynomial smoothing. The proposed method results in a novel estimator for the number of mixture components and is able to select active brain regions which correspond to active variables in a high-dimensional dynamic linear model.

E1339: **DeepTune: Visualization and interpretation of deep-network-based models in neuroscience**
Presenter: **Yuansi Chen**, ETH Zurich, Switzerland

Deep neural network models have been shown effective recently in predicting single neuron responses in primate visual cortex area V4. V4 is a midtier visual cortical area in the ventral visual pathway. Its functional role is not yet well understood. Despite the high predictive performance of deep neural network models, these models are generally difficult to interpret. This limits the applicability of these models in characterizing V4 neuron function. We propose the DeepTune framework as a way to elicit interpretations of deep neural network-based models of single neurons in area V4. Using a dataset of recordings of 71 V4 neurons stimulated with thousands of static natural images, we built an ensemble of 18 neural network-based models per neuron to accurately predict its response given a stimulus image. To interpret and visualize these models, we used a stability criterion to form optimal stimuli (DeepTune images) by pooling the 18 models together. These DeepTune images not only confirm previous findings on the presence of diverse shape and texture tuning in area V4, but also provide concrete and naturalistic visualization of predicted optimal stimuli of individual V4 neurons.

E1970: **Constrained functional additive models for estimating interactions between a treatment and functional covariates**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Hyung Park, Eva Petkova, Thaddeus Tarpey

A novel functional additive model is proposed which is uniquely modified and constrained to model nonlinear interactions between a treatment indicator and a potentially large number of functional/scalar covariates. We generalize functional additive regression models by incorporating treatment-specific components into additive effect components. A structural constraint is imposed on the treatment-specific components, to give a class of orthogonal main and interaction effect additive models. If primary interest is in interactions, we can avoid estimating main effects, obviating the need to specify their form and thereby avoiding the issue of model misspecification. The methods are illustrated with data from a clinical trial with imaging data as predictors.

EO132 Room MAL 253 ADVANCED STATISTICAL MODELLING AND APPLICATIONS Chair: Rosaria Simone

E0804: **Dissimilarity measure for ranking data via copula**
Presenter: **Silvia Angela Osmetti**, Universita Cattolica di Milano, Italy
Co-authors: Andrea Bonanomi, Marta Nai Ruscone

A new distance measure is defined for ranking data by using copula functions. This distance evaluates the dissimilarity between subjects expressing their preferences by rankings in order to segment them by hierarchical cluster analysis. The proposed distance builds upon the Spearmans grade correlation coefficient on a transformation of the ranks denoting the levels of the importance assigned by subjects under classification to k objects. The copula is a flexible way to model different types of dependence structures in the data and to consider different situations in the classification process. For example, by using copulae with lower and upper tail dependence, we emphasize the agreement on extreme ranks, when they are considered more important.

E0832: **Bilinear models for score building: When they should be used**
Presenter: **Brian Francis**, Lancaster University, United Kingdom
Co-authors: Elouise Davies

The focus is on the utility of bilinear models for score building in contingency tables and contrasts it with the correspondence analysis approach. The groundwork for using bilinear models for score building was laid time ago, and a set of rules for the instrumental variable against which the target variable is classified has been previously specified. Typical bilinear models used for this purpose include the log-multiplicative model and the correspondence analysis model. While this approach seems at first sight to be promising, there are issues relating to empty cells and sample size which often mean that the model fails to form exactly as intended. We discuss whether the mentioned rules need extending and determine whether similar rules are needed for correspondence analysis. An example is used from the problem of scaling crime harm and impact from survey data.

E0917: **Relationship between Kendall's tau and Goodman and Kruskal's gamma after ordinalizing a bivariate normal random variable**
Presenter: **Alessandro Barbiero**, Universita degli Studi di Milano, Italy

Goodman and Kruskal's gamma is a measure of association between two ordinal variables, based on probabilities of concordance and discordance, which can be seen as an adjustment to the discrete case of Kendall's rank correlation tau between two continuous random variables. By considering a standard bivariate normal random variable acting as a latent underlying distribution, we examine the relationship between its value of Kendall's tau (which is related to Pearson's correlation coefficient through a well-known analytic formula) and the value of gamma for a bivariate ordinalized distribution, by varying its margins and examining in particular uniform, unimodal symmetric, and triangular distributions. Based on this study, a procedure for finding the value of tau inducing a target value of gamma after ordinalization of a standard bivariate normal random variable is devised.

E0978: **Statistical solutions to improving bike-sharing systems**
Presenter: **Silvia Salini**, University of Milan, Italy
Co-authors: Giancarlo Manzi

Urban mobility is receiving increasing attention as one of the most important dimensions of the so-called smart city. If mobility must be sustainable, i.e., if it contributes to the improvement of quality of life, bike sharing can be viewed as a possible bridge between wellbeing and economic development. Understanding the mechanisms leading to a successful bike-sharing system (BSS) is indeed a hard task because of the many factors to be considered as, for example, the shape of the docking station network or the number of bikes deployed. One of the most important quandaries is in rightly predicting bike users behaviour, avoiding an uneven bike distribution among docking stations. We implement a decision framework to help policy makers to obtain optimal predictions of bike usage in the BSS BikeMi in Milan, Italy, using data on each bike itinerary from June 2015 to May 2018, including user and bike ID, check-in and check-out time and location, docking station availability, ride length and check-in and check-out time. We also use meteorological and environmental indicators. By using machine learning methods and Bayesian networks we model check out times and rental duration in order to better understand the bike users behaviour.

EG770 Room MAL 251 CONTRIBUTIONS IN LONGITUDINAL DATA ANALYSIS Chair: Mahmoud Torabi

E1757: **Joint modelling of longitudinal data involving time-varying covariates**
Presenter: **Reza Drikvandi**, Manchester Metropolitan University, United Kingdom

Longitudinal studies often produce data with both time-invariant and time-varying covariates. There are several major challenges with analysing such longitudinal data. For example, similar to classical regression models, standard models for longitudinal data ignore the covariate process and treat all covariates as fixed variables, while like the response variable, time-varying covariates also change over time and their evolution over time could provide important information. Also, longitudinal models assume that the response variable measured at each follow-up time depends on the time-varying covariates measured at that time point only, but the response variable could also depend on the previous measurements of time-varying covariates. We introduce a novel joint mixed model for the longitudinal outcome and the time-varying covariates to overcome such challenges. We use random effects to account for the association between the response and the time-varying covariates. We also incorporate P-spline functions into the joint model to capture the evolutions of the longitudinal response and the time-varying covariates over time. The proposed method is investigated theoretically and practically, and motivated by data from an AIDS cohort study in which HIV+ patients have CD4 cell count and viral load measured at repeated visits before and after receiving treatment.

E1936: **Deviations from normality: Effects on growth curve models**
Presenter: **Catarina Marques**, Instituto Universitario de Lisboa (ISCTE-IUL) and Business Research Unit (BRU-IUL), Portugal
Co-authors: Maria de Fatima Salgueiro, Paula Vicente

Latent growth curve models (LGCM) became recently a very popular technique for longitudinal data analysis: they allow individuals to have distinct growth trajectories over time. These patterns of change are summarized in relatively few parameters: the means and variances of the random effects (random intercept and slope), as well as the covariance between intercept and slope. Although the specified model structure imposes normality assumptions, the data analyst often faces data deviations from normality, implying mild, moderate or even severe values for skewness and/or kurtosis. The aim is to investigate the effect of data deviations from normality on the goodness of fit measures in LGCM. Using the VITA method to obtain data generating non-normal distributions, a Monte Carlo simulation study was conducted in order to assess the effects on the values of goodness of fit indices. LGCM with unconditional linear growth are considered. Three time points, and sample sizes ranging from 50 to 1000 observations are used. The impacts of such deviations on fit measures are discussed.

E1920: **Joint longitudinal models with informative time measurements**
Presenter: **Ines Sousa**, Minho University, Portugal
Co-authors: Adriana Vieira

In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in the worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristic should allow for an association between longitudinal and time measurements processes. We consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. The estimation of model parameters is addressed through maximum likelihood. We conduct a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In this case the follow-up time process should be considered dependent on the longitudinal outcome process.

E1705: **Automatic multivariate functional clustering for spatial longitudinal data**
Presenter: **Noritoshi Arai**, Chuo University, Japan
Co-authors: Toshihiro Misumi, Hidetoshi Matsui, Yoshihiko Maesono, Sadanori Konishi

Huge amount of multivariate longitudinal data with spatial information have been collected in recent years. So far, a problem of clustering for such data is little discussed. Functional clustering approach is one of the promising approaches for the multivariate longitudinal data. However, existing functional clustering does not consider the spatial dependence on the clustering algorithm even if the data contains spatial information. We introduce a novel multivariate functional clustering for spatial longitudinal data. A functional kriging approach with regularized basis expansions is applied to incorporate the spatial correlation for predicting functional data on unobserved point in the modeling process. After obtaining the predicted spatial functional data, a x-means clustering is implemented to the estimated coefficient vector of basis functions. The x-means is an extended method of k-means that automatically provides the optimal number of clustering. Numerical examples are presented to examine the effectiveness of our proposed clustering procedure.

CO628 Room MAL B02 PREDICTABILITY OF ASSET RETURNS Chair: Alexandros Kostakis

C0295: **Testing return predictability with the dividend-growth equation: An anatomy of the dog**
Presenter: **Erik Hjalmarsson**, University of Gothenburg, Sweden

Co-authors: Tamas Kiss

The dividend-growth based test of return predictability is similar to a likelihood-based test of the standard return-predictability model, treating the autoregressive parameter of the dividend-price ratio as known. In comparison to standard OLS-based inference, both tests achieve power gains from a strong use of the exact value postulated for the autoregressive parameter. When compared to the likelihood-based test, there are no power advantages for the dividend-growth based test. In common implementations, with the autoregressive parameter set equal to the corresponding OLS estimate, Cochrane's test also suffers from severe size distortions.

C1324: **Oil and equity return predictability: The importance of dissecting oil price changes**
Presenter: **Georgios Skoulakis**, University of British Columbia, Canada
Co-authors: Jinming Xue, Haibo Jiang

Oil price changes are documented no longer to predict G7 country equity index returns, in contrast to evidence based on data until the mid 2000s. Using a structural VAR approach, we decompose oil price changes into oil supply shocks, global demand shocks, and oil-specific demand shocks. The hypothesis that oil supply shocks and oil-specific demand shocks (global demand shocks) predict equity returns with a negative (positive) slope is supported by the empirical evidence over the 1986-2015 period. The results are statistically and economically significant and do not appear to be consistent with time-varying risk premia.

C1453: **Taking stock of long-horizon predictability tests: Are factor returns predictable?**
Presenter: **Michalis Stamatogiannis**, University of Liverpool Management School, United Kingdom
Co-authors: Tassos Magdalinos, Alexandros Kostakis

A critical assessment of long-horizon return predictability tests is provided by using highly persistent regressors. We show that the most commonly used test statistics are typically oversized, leading to spurious inference. As a remedy, we propose a simple Wald statistic, which can accommodate multiple predictors, exhibits excellent finite-sample properties regardless of the length of the predictive horizon, and is robust to the (unobservable) exact time series properties of the employed predictor(s). Employing this test statistic and a small set of variables that have been commonly used as proxies for business cycle conditions, we find evidence of predictability for "old" and "new" pricing factors with monthly returns. However, this evidence becomes weaker, not stronger, as the predictive horizon increases and disappears for most of the factors with annual returns. Overall, we question the incremental value of using long-horizon predictive regressions.

C1484: **What matters when: Time-varying sparsity in expected returns**
Presenter: **Andrea Tamoni**, Rutgers Business School, United States
Co-authors: Daniele Bianchi, Matthias Buechner

A measure of sparsity for expected returns within the context of classical factor models is provided. This measure is inversely related to the percentage of active predictors. Empirically, sparsity varies over time and displays an apparent countercyclical behavior. Proxies for financial conditions and for liquidity supply are key determinants of the variability in sparsity. Deteriorating financial conditions and illiquid times are associated with an increase in the number of characteristics that are useful to predict anomaly returns (i.e., the forecasting model becomes more dense). Looking at specific categories of characteristics, we find that variables classified as trading frictions are robustly present throughout the sample. A substantial amount of the time-variation in sparsity is attributable to the value, profitability, and investment categories. A strategy that exploits the dynamics of sparsity to time factors delivers substantial economic gain out-of-sample relative to both a random walk and a model based on preselected, well-know characeristics like size, momentum and book-to-market.

CO452 Room MAL B04 ADVANCES IN CREDIT RISK MODELLING II Chair: Raffaella Calabrese

C0827: **Network models to assess credit risk contagion**
Presenter: **Arianna Agosto**, University of Pavia, Italy
Co-authors: Paolo Giudici

A network model is proposed that decomposes credit risk contagion into a global component, made up of inter-sector contagion effects, and a local component made up of inter-institutional linkages. The model is effectively applied to a database containing time series of daily CDS spreads of major European companies and shows the importance of monitoring both global and local channels to assess credit risk contagion. The empirical application reveals evidence of a high inter-sector and inter-institutional vulnerability during the global financial crisis and during the European sovereign crisis. Our findings show that variations in sectorial risk factors play a significant role in the creation of vulnerable environments for risk propagation.

C1151: **On stress testing of credit default**
Presenter: **Jonathan Crook**, University of Edinburgh, United Kingdom
Co-authors: Viani Djeundje

The Bank of England requires large UK banks to undertake an annual stress test of their regulatory capital. The ECB and the Federal Reserve Board require periodic stress tests and stress testing is a requirement under the Basel Accords. The usual procedure is for the central bank to provide a macroeconomic scenario and for each bank to predict the amount of capital it would be required to hold for credit, market and operational risk to protect depositors against plausible but unexpected adverse events. Much of the literature considers methods to predict the capital required at a portfolio level for credit risk losses. However uncertainty over the state of the macroeconomy is only one source of risk when predicting the amount of capital needed for credit risk. There is also a growing literature on model risk and the literature on stress testing usually omits this source of uncertainty. We consider sources of uncertainty other than macroeconomic risk. We estimate hazard functions for the probability of default using a large sample of credit card accounts. We estimate the separate impact of mis-estimation risk, volatility risk and behavioural risk on the value at risk and expected shortfall and so on the amount of capital needed to protect depositors. We find that each type of risk is relatively modest. The work differs from the literature by undertaking an account level analysis and by estimating the relative magnitudes of different types of model risk.

C1242: **Use of social media big data for predicting credit rating changes of companies**
Presenter: **Leonie Tabea Goldmann**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Jonathan Crook

Predicting the financial performance of companies using various data sources has been the focus of a large number of studies. However, current models can still be improved to achieve even better predictions. Several contributions are made. First, a new method is introduced to more accurately predict the probability that a companys credit rating changes. More specifically, the vast increase in social media data put out by companies is exploited which enables to develop more predictive models than those developed in the past. By analysing the tweets put out by different companies, we show that there is a correlation between specific words in the tweets and the credit rating. This is done by using differential language analysis, a method which is usually used in psychology and health studies and has not been used in a financial context before. Second, we will show that using these words in a predictive model with additional twitter variables, such as the tweet frequency, sentiment score and length leads to an increase in predictive power when comparing with a model not containing these predictors.

C1678: **Alternative methods of default risk estimation**
Presenter: **Patrycja Chodnicka - Jaworska**, University of Warsaw, Poland

The basic goal is to analyse macroeconomic and financial factors influencing the default risk by taking into account the business line and credit

ratings. A research question has been put as follows: Are credit ratings estimated by the same methods for analysing the default risk, by taking into account the business lines? Two hypotheses are considered. The first one is: Countries risk has a significant influence on credit ratings changes. The second one is: The determinants of credit ratings assigned by major rating agencies are differentiated by taking the type of the business lines. For verification of these hypotheses the quarterly data form the Thomson Reuters database were collected. As dependent variables, the long term issuer credit ratings proposed by the recognizable CRAs from 1990 to 2017 period of time are used. The analysis has been prepared in the sub-samples according to: the type of credit rating, the domestic and foreign notes and the type of business lines.

CO735 Room MAL B18 IDENTIFICATION IN SVARS Chair: Robin Braun

C0938: **The macroeconomic effects of oil supply news: Evidence from OPEC announcements**
Presenter: **Diego Kaenzig**, London Business School, United Kingdom

A novel approach is proposed to study the macroeconomic effects of oil prices, exploiting institutional features of OPEC and high-frequency data. Using variation in futures prices around OPEC announcements as an instrument, we identify an oil supply news shock. These shocks have statistically and economically significant effects. Negative news leads to an immediate increase in oil prices, a gradual fall in oil production and an increase in inventories. This has consequences for the U.S. economy: activity falls, prices and inflation expectations rise, and the dollar depreciates - providing evidence for a strong channel operating through supply expectations.

C0935: **The role of inflation expectations in the transmission of monetary policy**
Presenter: **Catalina Martinez Hernandez**, Freie Universitaet Berlin/ DIW Berlin, Germany

The role of inflation expectations for explaining heterogeneous monetary transmission in several countries of the Euro Area is empirically investigated. We analyze if imperfect information in the expectations-formation process explain the differences in the responses of inflation to conventional and unconventional monetary policy shocks. Since the number of variables that the ECB monitors is large, we concentrate on a data-rich environment framework and consider a Large Bayesian Proxy Vector Autoregression. We estimate this model by combining two strands of the literature on Bayesian Macroeconometrics: The Large Bayesian VAR and the Bayesian Proxy VAR. We find evidence that incomplete information in the formation of expectations could be considered as an explanation for the different responses of inflation to monetary policy in these countries. The policy implications are crucial for understanding the role of inflation expectations for monetary transmission especially in periods of unconventional tools.

C0280: **Identifying smooth transition vector autoregressivemodels**
Presenter: **Martin Bruns**, University of East Anglia, United Kingdom

We develop a way of identifying Smooth Transition Vector Autoregressive models using sign restrictions. In so doing, we offer an alternative to the recursive identification typically used for these models. What makes sign restrictions tractable despite the nonlinearities of the model is the use of a Bayesian approach. We then use the model and argue that monetary policy shocks have a smaller effect on output when the economy is experiencing relatively high levels of uncertainty. The results cannot be driven by the higher price flexibility documented by the literature in periods when uncertainty is high. In fact, we find that not only output but also prices display a limited response to monetary shocks when uncertainty is high. The results suggest that the effectiveness of monetary policy is driven by mechanisms that go beyond price stickiness.

C0866: **The importance of supply and demand for oil prices: Evidence from identification by non-Gaussianity**
Presenter: **Robin Braun**, Bank of England, United Kingdom

New evidence is provided on the relative importance of supply and demand shocks for fluctuations in oil prices. To estimate their effects, a structural vector autoregressive (SVAR) model for the global oil market is identified by non-Gaussianity. The identification approach is further refined by ruling out economically unreasonable oil price elasticities a priori. To incorporate this prior information in a coherent way, a new Bayesian SVAR is developed where the unknown distributions of the structural shocks are modeled nonparametrically. The empirical findings indicate that oil supply shocks have been minor drivers of oil prices post 1985. In terms of contributions to the long term forecast error variance of oil prices, the model arrives at median estimates between 1% and 13% depending on the exact prior specifications.

CO713 Room MAL B35 PANEL DATA METHODS FOR INTEGRATED SERIES Chair: Anindya Banerjee

C1149: **Partially heterogeneous tests for Granger non-causality in panel data**
Presenter: **Yiannis Karavias**, University of Birmingham, United Kingdom
Co-authors: Arturas Juodis

The power of Granger non-causality tests in panel data depends on the type of the alternative hypothesis: feedback from other variables might be homogeneous, homogeneous within groups or heterogeneous across different panel units. Existing tests have power against only one of these alternatives and may fail to reject the null hypothesis if the specified type of alternative is incorrect. A new Union-Intersections (UI) test is proposed which has correct size and good power against any type of alternative. The UI test is based on an existing test which is powerful against heterogeneous alternatives and a new Wald-type test which is powerful against homogeneous alternatives. The Wald test is designed to have good size and power properties for moderate to large time series dimensions and is based on a bias-corrected split panel jackknife-type estimator. Evidence from simulations confirm the new UI tests provide power against any direction of the alternative.

C1292: **Bounds testing for panel unit roots using common correlated effects estimators**
Presenter: **Josep Lluis Carrion-i-Silvestre**, Universitat de Barcelona, Spain
Co-authors: Anindya Banerjee

A panel data unit root statistic is proposed which has cross-section dependence driven by unobserved common factors that are approximated by means of the common correlated effects estimation method. The null hypothesis of panel data unit root focuses on the idiosyncratic component, although the statistical inference is conducted using a bounds testing strategy. Proceeding in this way the analysis takes into account that the cross-section dependence might be driven by $I(1)$ non-stationary common factors, $I(0)$ common factors, or a mixture of both $I(1)$ and $I(0)$ common factors and, therefore, extending some existing proposals in the literature.

C0235: **Consumption, income, wealth and 10-year treasury rate in the 51 States of US: A panel cointegration approach**
Presenter: **Dimitra Kontana**, University of Macedonia, Greece
Co-authors: Stylianos Fountas

The long-run relationship between consumption, financial and housing wealth is investigated by considering the 10-year treasury rate for the 51 States of US. By using quarterly observed up-date data from 1975 (1) to 2018 (1) we conduct two kinds of tests in order to perform the panel cointegration analysis: a residual based cointegration test for cross-sectionally independent panel and a panel cointegration test for cross-sectional dependence panel. All the test statistics reject the null of no cointegration hypothesis at 1 percent level of significance. The empirical findings show that housing wealth elasticity of consumption ranges from 0.072 to 0.115 percent, while the financial wealth elasticity of consumption is between 0.044 to 0.080 percent. The latter is less significant than the income and housing wealth according to the first statistic, and insignificant according to the second statistic. These results are in line with the findings of early researchers, who performed alternative to cointegration methods. Finally, Granger causality tests show that there is a bidirectional short-term causality between per capita consumption, income and financial wealth in the short run and between all the variables in the long run.

C1637: **Fully modified least square for multicointegrated systems**
Presenter: **Igor Kheifets**, ITAM, Mexico
Co-authors: Peter CB Phillips
Multicointegration is traditionally defined as a particular long run relationship among variables in a parametric vector autoregressive model. We depart from the parametric model. This allows us to provide the explicit relationship from which the multicointegration arises and reveal the leading role that the singularity of the long run conditional covariance matrix plays in determining multicointegration. Considering multicointegration in a semiparametric framework has an advantage that the short run dynamics of time series does not need to be modeled. We show that in a semiparametric triangular representation of cointegrated time series, multicointegration results in a singular long run covariance. We derive a convergence of fully modified regression estimator in case of singular long run covariance. We obtain faster rates of convergence along particular directions, these rates and asymptotic distribution depend on the conditional one-sided long run covariance estimator used after the first stage. We also show that in the presence of singularity the Wald test for restrictions on the regression coefficient has nonstandard distribution, depends on nuisance parameters and is conservative if restrictions isolate those directions and is invariant to singularity otherwise.

CO707 Room MAL B36 SENTOMETRICS Chair: Kris Boudt

C0377: **Conservative and timely ESG-compliant investment universe screening using textual sentiment analysis**
Presenter: **Samuel Borms**, Universite de Neuchatel, Switzerland
Co-authors: David Ardia, Kris Boudt, Andres Algaba
Combining textual sentiment analysis and econometric techniques, we improve the construction of an ESG best-of-class investment universe by more timely detection of companies and countries likely to succumb to a sustainability downgrade. Our approach consists of several layers. First, we score the sentiment in the relevant news articles employing various ESG-specific lexicons. Second, we aggregate the textual sentiment scores and remove the noise through application of the Kalman filter. Third, we use predictive modelling and forecast combination over multiple horizons to assess the likelihood of downgrading for every investable stock. Fourth, we define decision rules to construct and update both the sustainable investment universe and a blacklist. In an empirical portfolio application, we test for the added value of our more accurate and timelier sustainability signals.

C0663: **The GWP R package: Generalized word power approach of calibrating lexicons**
Presenter: **Keven Bluteau**, Institute of Financial Analysis, University of Neuchatel, Switzerland
Text analysis tools are becoming increasingly important in financial and economic research as recent research projects focus on textual analysis to answer new hypotheses or increase forecasting power. The standard approach to extract sentiment from textual data is the bag-of-word method, where a list of words and corresponding scores, referred to as a lexicon, are used to quantify textual data to uncover, for example, textual sentiment. This list of words and scores are often manually set up by the researcher in an ad-hoc way. We present the R package GWP which implements the Generalized Word Power approach, a data-driven lexicon calibration framework. The package provides several helper functions to organize textual data as well as a Shiny interface to help validating the calibrated lexicons.

C0842: **Consumer confidence media indexation**
Presenter: **Andres Algaba**, Vrije Universiteit Brussel, Belgium
Co-authors: Kris Boudt, Samuel Borms
The consumer confidence index is an important (leading) economic indicator in Belgium. Currently, it is measured by the National Bank of Belgium via a monthly survey. We argue that this survey method can be improved with the qualitative information embedded in news media articles. By using a standard lexicon approach, we extract daily sentiment scores from news articles of the national Belgian press agency. We then use these daily sentiment scores and monthly observations on the consumer confidence index in a mixed-frequency state space framework to provide nowcasts of consumer confidence index.

C2001: **Media bias detection using sentometrics**
Presenter: **Jeroen Van Pelt**, Vub, Belgium
Co-authors: Andres Algaba, Samuel Borms, Kris Boudt
Media bias can strongly influence the public opinion on important topics which in turn can ultimately affect peoples decision-making process. While the social sciences recognize the important impact of media bias, automated and scalable methods to detect textual bias in news articles are lacking. Media bias analysis tools are useful to quantify and visualize the relative difference between the coverage and reporting of a news fact by one news medium compared to its peers. We identify four dimensions of media bias through reporting, and define appropriate metrics for each dimension. These bias metrics can be constructed using a combination of sentiment analysis and econometrics, which we refer to as sentometrics. The reporting bias can then be computed by comparing these metrics to its peer articles that have been identified via a cosine similarity based matching tool. Finally, we empirically show the effectiveness of automated bias detection in media news articles.

CO422 Room Montague MODELLING, FORECASTING AND ACCURACY Chair: Mauro Costantini

C0552: **Forecasting commodity futures returns with stepwise regressions: The contribution of commodity-specific factors**
Presenter: **Manuela Pedio**, Bocconi University, Italy
Co-authors: Massimo Guidolin
The aim is to assess whether three well-known commodity-specific variables (basis, hedging pressure, and momentum) may improve the predictive power for commodity futures returns of models otherwise based on macroeconomic factors. We compute recursive, out-of-sample forecasts for the monthly returns of fifteen commodity futures, when the estimation is based on a stepwise model selection approach under a probability-weighted regime-switching regression that identifies different volatility regimes. We systematically compare these forecasts with those produced by a simple AR(1) model that we use as a benchmark and we find that the inclusion of commodity-specific factors does not improve the forecasting power. We perform a back testing exercise of a mean-variance investment strategy that exploits any predictability of the conditional risk premium of commodities, stocks, and bond returns, also consider transaction costs caused by portfolio rebalancing. The risk-adjusted performance of this strategy does not allow us to conclude that any forecasting approach outperforms the others. However, there is evidence that investment strategies based on commodity-specific predictors outperform the remaining strategies in the high-volatility state.

C0716: **Volatility forecasting and performance evaluation**
Presenter: **Eirini Bersimi**, University of Kent, United Kingdom
The aim is to compare and evaluate alternative univariate volatility forecasting specifications in terms of their out-of-sample performance. We consider a plethora of specifications ranging from non-parametric approaches, parametric models from the GARCH family under different distributional assumptions and Realized Volatility models to forecast combinations. To evaluate the forecasts, pairwise comparisons are performed using a previous test of equal predictive ability under alternative loss functions. Additionally, the Model Confidence Set (MCS) and the Superior Predictive Ability (SPA) tests are performed for three robust loss functions. An empirical application on the S&P 500 index shows that there is no single specification that outperforms the rest and that simple specifications such as the TGARCH and the GARCH model appear superior. Our findings are robust to the stock index included as our main conclusions are conformed for six additional considered stock indices.

C0834: **A forecast-based consumer sentiment index**
Presenter: **Claudio Lupi**, University of Molise, Italy
Co-authors: Giancarlo Bruno, Marco Centoni

In recent years there has been a growing interest in using consumer surveys as leading indicators of real economic activity, especially to assess the future path of private consumption. Consumer sentiment indices are widely used and analysed with this aim, but are generally produced from consumer surveys using rather ad hoc criteria and weights. However, consumer surveys amount to very large data sets whose value is probably still underrated. We offer a forecast-based alternative measure of the consumer sentiment with a well defined economic meaning. The indicator is based on the canonical correlations between consumption and finely disaggregated consumer survey data. The underlying idea is to exploit the property of canonical correlation in order to find the linear combination of the consumer survey items that maximises correlation with final consumption. The new indicator is computed on Italian consumer survey data and is derived for total final consumption expenditure of household and for consumption expenditure by durability. Some comparisons with the existing consumer sentiment indicator are derived and discussed.

C0667: **DGSE models with expectations correction: Misspecification, forecasting errors and directional accuracy**
Presenter: **Mauro Costantini**, University of L'aquila, Italy
Co-authors: Giovanni Angelini

The forecasting performance of small-scale New-Keynesian models with expectations correction is considered. These models are designed to reduce miss-specification. A comprehensive comparative forecasting evaluation of different specifications through a Monte Carlo analysis is offered, so to establish whether DSGE models with expectations correction perform better than other macro structural models. Further, an empirical application, based on frequentist estimation of DSGE models with expectations correction for the US data, is provided.

CO218 Room Woburn APPLIED MACROECONOMIC AND MACRO-FINANCIAL TOPICS II Chair: Wojtek Paczos

C0294: **Imperfect financial markets and the cyclicality of social spending**
Presenter: **Wojtek Paczos**, Cardiff University, United Kingdom
Co-authors: Maren Froemel

The link between default risk and the cyclicality of redistribution is explored. Empirically, we establish a stylized fact, that countries with higher sovereign risk have more procyclical fiscal policy. We build a small open economy model with heteregenous households, government consumption, social transfers, and endogenous default risk to rationalize this fact. With low default risk social spending is countercyclical, inequality is procyclical, and external debt is used to smooth distortionary taxation. With high default risk, social spending is procyclical and accounts for most of fiscal adjustment when risk premia are positive, because taxation becomes costly for the government.

C0297: **Comovement changes between stocks and bonds: Evidence from a class of large dimensional threshold group-factor models**
Presenter: **Daniele Massacci**, Kingś College London, United Kingdom
Co-authors: Mirco Rubin, Dario Ruzzi

The determinants of time-varying comovement between U. S. stock and government bond returns are studied as driven by regime changes in equity market volatility measured by the VIX. We introduce a novel group-factor model with regime-dependent factor loadings: the prevailing regime within each group of securities depends on the previous value of the VIX with respect to a threshold. For each security within a group, the large dimension of the panel allows to separately identify systematic and diversiable risk components. Within each regime, we measure comovement between the groups as the share of systematic correlation between them. Using 30 years of monthly data, we find that a lagged value of the VIX larger than 27 determines a significant increase in comovement between U.S. equity returns and the term structure of U.S. interest rates. We propose a test for the change in the average systematic correlation between the two panels in the two regimes.

C0305: **State-dependent pricing and its implications for monetary policy**
Presenter: **David Meenagh**, Cardiff University, United Kingdom

Strong evidence now exists both in macro and micro data that price/wage durations are dependent on the state of the economy and especially inflation. We embed this dependence in a macro model of the US that otherwise does well in matching the economy's behaviour in the last three decades; it now also matches it over the whole post-war period. This finding implies a major role for monetary policy in influencing the economy's price stickiness. We find that by targeting nominal GDP monetary policy can achieve high price stability while also preventing large cyclical output fluctuations.

C0446: **External adjustment with a common currency: The case of the euro area**
Presenter: **Alberto Fuertes**, Bank of Spain, Spain

The behaviour of the external adjustment path for the four main economies in the euro area is analyzed. We find a structural break in the behaviour of the net external position at the time of the introduction of the euro for France, Italy and Spain, pointing out that the inception of the common currency changed their external adjustment process. Germany does not show this structural break, being its external position more affected by other events such as the country reunification in 1989. We also find that France and Italy will adjust the net external position mainly through the valuation component of external adjustment, while Germany and Spain will restore their external balance mostly through the trade component. The common currency area exacerbated Germany's net creditor position as the evolution of the euro has reacted to the external adjustment needs of debtor countries such as Italy and Spain.

CO416 Room Chancellor's Hall BAYESIAN ECONOMETRICS Chair: Silvia Montagna

C1491: **Factor multivariate realized stochastic volatility model**
Presenter: **Yuta Yamauchi**, University of Tokyo, Japan
Co-authors: Yasuhiro Omori

Although modelling time-varying volatility and correlations of multivariate asset returns is one of the most important problems in the financial risk management, it has been difficult to obtain stable inference of covariance of asset returns due to the high dimensionality of parameters in dynamic covariance structure. One major solution to reduce the number of parameters is to introduce factor structure assuming that a small number of common factors describe the dynamics of time-varying covariance matrices as discussed in the factor stochastic volatility models. We propose parsimonious modelling of multivariate asset returns based on dynamic factor stochastic volatility models with leverage effect, using high-frequency data as additional observations. Firstly, to stabilize the estimation of time-varying parameters, we incorporate additional observations based on intraday asset returns and market indices. We use realized measures for covariance of asset returns based on intraday asset returns such as realized covariance, and latent factors of asset returns such as market indices. Secondly, we reduce the number of parameters of leverage effect, omitting the leverage effect between each asset and each volatility of asset. We only introduce leverage effect between each latent factor and each volatility of assets.

C1505: **Bayesian dynamic fused lasso**
Presenter: **Kaoru Irie**, University of Tokyo, Japan

The new class of Markov processes is proposed to model the flexible shrinkage effects in the prior of time-varying parameters for time series analysis. The transition density of the new process consists of two penalty functions, similar to Bayesian fused LASSO in its functional form, that

shrink the current state variable to its previous value and zero. The normalizing constant of this density, which is not ignorable in the posterior computation, is shown to be essentially the log-geometric mixture of double-exponential densities and treated as a part of the likelihood. By using this new process as a prior on dynamic coefficients, the dynamic regression model is conditionally Gaussian and linear in state variables, for which the posterior can be computed efficiently by the forward filtering and backward sampling in Gibbs sampler. The log-geometric distributed latent variable is understood as the amount of shrinkage to zero realized in the posterior and can be used to detect periods in which the corresponding dynamic coefficient becomes inactive. The new prior is compared with the standard double-exponential prior in the estimation of, and prediction by, the dynamic linear models with the simulated datasets for illustration. It is also applied to the time-varying vector autoregressive models for the US macroeconomic data, which exemplifies the use of the new prior as an alternative of the dynamic model of variable selection type, such as the latent threshold models.

C1580: **Predictive properties of forecast combination, ensemble methods, and Bayesian synthesis**
Presenter: **Kenichiro McAlinn**, Temple University, United States
Co-authors: Kosaku Takanashi

The aim is to study the theoretical predictive properties of multiple classes of forecast (model) combination strategies, motivated by a recent development in a foundational Bayesian framework called Bayesian predictive synthesis. A novel strategy based on continuous time stochastic processes is developed, where the predictive error processes are expressed as stochastic differential equations that are evaluated using Itos lemma. As a result, we show that a class of Bayesian predictive synthesis functions, which we categorize as non-linear synthesis, entails an extra term in the stochastic process that is interpreted as a shrinkage term on the error process, effectively improving forecasts. Theoretical properties are examined and shown that this subclass improves expected squared forecast error over any and all linear combination, averaging, and ensemble of forecasts. A finite sample simulation study is presented to illustrate our results. The results imply directions for further research and inquiry, which are discussed.

C1596: **The state of industry statistics in online AB experiments**
Presenter: **Michael Lindon**, Optimizely, United States

A survey of the most popular statistical methodologies used in the experimentation and growth industry concerning online A/B experiments is provided. The online experimentation space concerns running numerous web based experiments on millions of users to measure the effect of various treatments in increasing metrics of interest - user signups, revenue spend per user and user conversions for example. Unfortunately statistical best practices are often overlooked in this domain, for which the usual offenders are incorrectly adjusting for multiple testing and continuous monitoring. The literature for sequential hypothesis testing is reviewed, its application to online experimentation, and some very recent developments in extending these methodologies to combinations of composite hypotheses are discussed. A juxtaposition of Bayesian, frequentist and conditional frequentist approaches to this problem are presented throughout, with a discussion of their commonalities and differences.

CO458 Room Court MIXTURE MODELS IN ECONOMETRICS Chair: Ralf Wilke

C0492: **Modeling frailty correlated defaults with multivariate latent factors**
Presenter: **Benjamin Christoffersen**, Copenhagen Business School, Denmark
Co-authors: Rastin Matin

Firm-level default models are important for bottom-up modeling of the default risk of corporate debt portfolios. However, models in the literature typically have several strict assumptions which may yield biased results, notably a linear effect of covariates on the log-hazard scale, no interactions, and the assumption of a single additive latent factor on the log-hazard scale. Using a sample of US corporate firms, we provide evidence that these assumptions are too strict and matter in practice and, most importantly, we provide evidence of a time-varying effect of the relative firm size. We propose a frailty model to account for such effects that, unlike previous models in literature, can provide forecasts for arbitrary portfolios as well. The proposed model displays superior out-of-sample ranking of firms by their default risk and forecasts of the industry-wide default rate during the recent global financial crisis.

C0565: **Addressing measurement error in the estimation of labor market transitions**
Presenter: **Daniel Borowczyk-Martins**, Copenhagen Business School, Denmark

A new approach is developed to estimate transition probabilities based on a series of repeated cross sections and when the variable recording individuals' past state is exposed to classification and recall errors. The problem and the solution are motivated by the estimation of the canonical model of worker flows using the International Labor Organization classifications of labor market states and microdata from the European Union Labor Force Survey. We specify the data-generating process of observed individual-level transition probabilities as a mixture of two multinomial logit models (one for the unobserved transition probabilities and another for the measurement-error process) and inform its estimation with two sources of auxiliary data.

C1160: **Data driven estimation of group structures in individual fixed effects models**
Presenter: **Kristina Maria Zapp**, ZEW Mannheim, Germany
Co-authors: Ralf Wilke

A linear fixed effects estimation is combined with clustering techniques to reduce the dimension of the fixed effects in a data driven way. We assume that individual fixed effects follow an unobserved grouped pattern and the size of the unobserved effects can be correlated with observable characteristics. The dimension reduction allows for the inclusion of individual time invariant characteristics in the fixed effects estimation. We apply density based clustering methods with endogenous computation of the number of groups and compare it to prototype methods (k-means) with given number of groups. We study the performance of the combined estimator in a Monte Carlo simulation that is to a large extend based on administrative labor market data from Germany.

C1769: **Statistical inference for mixture GARCH models**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia - Dipartimento di Economia Marco Biagi, Italy

Mixture generalized autoregressive conditional heteroskedastic models are considered. A new iteration algorithm of type EM is proposed for the estimation of model parameters. The maximum likelihood estimates are shown to be consistent, and their asymptotic properties are investigated. More precisely, we derive simple expressions in closed form for the asymptotic covariance matrix and the expected Fisher information matrix of the ML estimator. Finally, we study the model selection and propose testing procedures. Applications and examples illustrate the results.

CO424 Room SH349 ECONOMETRIC METHODS FOR SPORT DATA MODELLING AND FORECASTING Chair: Luca De Angelis

C0214: **Information, prices and efficiency in an online betting market**
Presenter: **Carl Singleton**, University of Reading, United Kingdom
Co-authors: Guy Elaad, James Reade

The odds (or prices) set by fifty-one online bookmakers for the result outcomes in over 16,000 association football matches in England since 2010 are studied. Adapting a methodology typically used to evaluate forecast efficiency, we test the Efficient Market Hypothesis in this context. At the overall market level, we found no statistically significant evidence which could reject an efficient market hypothesis of the online betting market for English football match results. The odds offered by online bookmakers were generally not biased towards any particular result outcome, nor did

they feature the favourite-longshot bias, which has been documented in other betting markets. But individual bookmaker-specific markets are not efficient, since they fail to use the information contained in their competitors' odds. There is also suggestive evidence that the increased competition facing online bookmakers has reduced commission rates and profit margins.

C0501: **Market prices reaction to information: A quasi-experiment**

Presenter: **Romain Gauriot**, New York University Abu Dhabi, United Arab Emirates

Co-authors: Lionel Page

The aim is to investigate whether prices on a financial market over or underreact to new information. To do so, we use a quasi-experimental setting where we can compare the arrival of an informational shock to a very similar situation where this shock did not happen. Specifically, we look at in-play betting market prices for Association Football matches and we investigate how their prices react when a shot lands on a post. We use shots hitting the post and landing inside the goal as the events of interest (informational shock) and the shots bouncing off the post as the counterfactual situations. Comparing these two situations, we find that most of the time, prices react efficiently to the arrival of a goal. However, we find under-reaction in some cases where the informational shock is large. In particular, in the last 20 minutes of the match there is a large under-reaction when the score moves from a draw to a winning situation. These results are robust to adjustment for multiple testing.

C1172: **Informational efficiency and price reactions in exchange betting markets**

Presenter: **Luca De Angelis**, University of Bologna, Italy

Co-authors: Giovanni Angelini

The degree of efficiency of exchange betting markets is investigated. In particular, we test whether prices on the exchange market are set efficiently both before the beginning of a football match and in the aftermath of an event occurred during the match. First, we propose a forecast-based approach for formally testing the efficiency of pre-match prices on the exchange betting markets. Then, using event study analysis on high-frequency data, we examine the reaction of prices to events and the arrival of major news. In particular, we analyse the post-event evolution of in-play odds and their dynamic behaviour at different horizons. We test for informational efficiency in football exchange betting markets in three different directions: (i) we investigate whether there is evidence of mispricing in in-play odds as reaction to major news events (i.e. goals, red cards), (ii) we test whether the arrival of new information on the market creates systematic bias which can be exploited to set a profitable betting strategy, (iii) we focus on possible over/under-reaction by investigating the main drivers which may create deviations from efficiency. To do so, we consider a large dataset comprising prices collected every ten seconds from Betfair Exchange for all the English Premiership matches played in the five seasons from 2009/10 to 2013/14.

C1300: **Forecasting football match outcomes with big data and bigger methods**

Presenter: **James Reade**, University of Reading, United Kingdom

Co-authors: Shixuan Wang, Luca De Angelis, Carl Singleton

A range of methods are considered for forecasting outcomes of football matches. In particular, we apply big data, and econometrics methods designed to cope with a large range of explanatory variables (LASSO, general-to-specific) where possible to forecast outcomes. We evaluate methods using a range of betting-related strategies. We apply these methods to football matches across numerous leagues in Europe, including the big five Leagues of England, France, Germany, Italy, and Spain.

Saturday 14.12.2019 16:45 - 18:50 Parallel Session F – CFE-CMStatistics

EO080 Room CLO B01 STATISTICS FOR HILBERT SPACES Chair: Gil Gonzalez-Rodriguez

E0901: **A classification tree for functional data**
Presenter: **Jan Gertheiss**, Helmut Schmidt University, Germany
Co-authors: Annette Moeller

Many standard tools for data analysis already have their functional counterparts tailored to the specific properties of functional data. We introduce a novel classification tree specifically designed to deal with functional predictors. Partitioning for a chosen predictor in a specific node of the tree is based on comparing each observational curve in that node to the class-specific mean in terms of a (functional) distance measure. A curve under consideration is assigned to the class to whose mean curve it is closest in terms of the chosen metric.

E0950: **Varying-coefficient functional additive models**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan

Varying-coefficient functional linear models consider the relationship between a scalar response and a functional predictor, where the coefficient function depends on another variable. It then accounts for the influence of the variable to the response. We extend the varying-coefficient functional linear model to the framework of an additive model and propose a varying-coefficient functional additive model. It represents the nonlinear relationship between the response and the predictor by introducing nonlinear functions of functional predictor. We estimate the varying-coefficient functional additive model by the penalized likelihood method. The effectiveness of the proposed model is investigated through simulation and real data analysis.

E1064: **Spectral analysis of high-dimensional functional time series**
Presenter: **Yirui Liu**, London School of Economics, United Kingdom
Co-authors: Xinghao Qiao

Statistical modeling for high-dimensional functional time series has attracted increasing attention in recent years. Under such a scenario, not only the number of functional variables, p, is large compared to the number of functional observations, n, but each function itself is infinite-dimensional with temporal dependence across observations. A useful approach to handle multivariate stationary functional times series by estimating its spectral density matrix function based on the averaged periodogram. We present a non-asymptotic theory for such spectral analysis. In particular, we derive useful concentration bounds on estimated spectral density matrix functions, which serves as a fundamental tool for further consistency analysis in large p, small n settings. We illustrate the usefulness of our derived concentration results in two examples. The first example considers the regularized estimation of high-dimensional spectral density matrix functions via functional thresholding. The second example considers a dynamic functional principal component analysis (FPCA), which is the main dimension reduction technique to handle functional time series. We rely on our developed concentration results to investigate the consistency properties of estimated terms for both examples under high dimensional scaling $\log(p/n)$ goes to zero.

E1294: **Depth analysis for sparse functional data**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States

Functional data analysis is an exciting developing area in statistics where the basic unit of observation is a function. Many different statistical methods, such as principal components and regression, have been extended to functional data. In the last decade there has been an intensive development of different notions of data depth for functional data which have been proven to be a powerful robust nonparametric tool for analyzing functions. In general, a data depth is a function that measures the centrality of an observation within a population or sample. It provides a rigorous way of ranking observations from center-outward and of defining robust statistics such as medians and trimmed means. The notions of depth for functional data introduced in the literature are designed for sample of curves that are measured on a common and dense grid. In practice, curves are often observed at subject-dependent and sparse grids and therefore, they have to be first estimated in a common grid. Standard functional depth analysis has ignored the inherent uncertainty associated with the preliminary curve estimation step. We design a general procedure that allows the analysis of depth to explicitly address sparsity and to take curve uncertainty estimation into account. In a simulation study we show the performance of the proposed approach in different settings changing the types of sparsity of the simulated curves.

E1824: **Local dimension reduction using small ball probability factorization**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Jean-Baptiste Aubin

The small–ball probability of a Hilbert valued process is considered. Recent works on its factorization put the lights on a factor that is used to study the local dimension of a Hilbert valued process. The properties of this factor are studied, and an estimator is introduced. It turns out that such estimator is consistent and asymptotically normal distributed. Features of such estimator provide insight on the local dimensionality of the process.

EO460 Room MAL B02 NOVEL APPROACHES FOR HIGH-DIMENSIONAL MEDICAL DATA Chair: Damian Brzyski

E1577: **Sensitivity biplots**
Presenter: **Julia Fukuyama**, Indiana University, United States

In high-dimensional biological datasets, distances customized to the biological system being studied are often used in conjunction with multi-dimensional scaling in order to obtain a lower-dimensional representation of the data. This low-dimensional representation can be used either for exploratory analysis or to obtain a smaller set of variables for later statistical analysis. Multi-dimensional scaling has an advantage over linear methods such as PCA in that it can be used to embed any set of distances between samples, but it has the disadvantage of being difficult to interpret in terms of the original variables. To overcome this limitation, we present a generalization of PCA biplots to multi-dimensional scaling. These plots describe the relationship between the multi-dimensional scaling embedding space and the original variables, allowing a visualization of variable importance in different parts of the space. We illustrate the method on a human microbiome dataset, showing how it gives insight into both the distance used to construct the embedding and the relevant biology.

E1627: **The adaptive incorporation of multiple sources of information in brain imaging via penalized optimization**
Presenter: **Damian Brzyski**, Wroclaw University of Science and Technology, Poland

The use of multiple sources of information in regression modeling has recently received a lot of attention in the statistical and brain imaging literature. A novel, fully-automatic statistical procedure is introduced which addresses the problem of linear regression coefficients estimation in the situation when the additional information about connectivities between variables is given. Our method, Adaptive Information Merging Estimator for Regression (AIMER) enables for the incorporation of multiple sources of such information as well as for the division of one source into pieces and determining their impact on the estimates. We performed extensive simulations to visualize the desired adjusting properties of our method and show its advantages over the existing approaches. We also applied AIMER to analyze structural brain imaging data and to reveal the association between cortical thickness and HIV-related outcomes.

E1628: **Adaptive Bayesian SLOPE: High-dimensional model selection with missing values**
Presenter: **Malgorzata Bogdan**, University of Wroclaw, Poland

Co-authors: Wei Jiang, Julie Josse, Blazej Miasojedow, Veronika Rockova

The selection of variables with high-dimensional and missing data is a major challenge and very few methods are available to solve this problem. We propose a method – adaptive Bayesian SLOPE – which is an extension of the Sorted L-One Penalized Estimator within a Bayesian framework and which allows us to simultaneously estimate the parameters and select variables for FDR control for large data despite missing values. Extensive simulations highlight the good behavior in terms of power, FDR and estimation bias under a wide range of simulation scenarios. Finally, we consider an application for prediction of the level of platelets for severely traumatized patients from Paris hospitals. We demonstrate that beyond the advantage of selecting relevant variables, which is crucial for interpretation, ABSLOPE has excellent predictive capabilities. The methodology is implemented in the R package ABSLOPE, which incorporates C++ code to improve the efficiency of the proposed method.

E1630: **Uniqueness and model selection for SLOPE estimator**

Presenter: **Patrick Tardivel**, University of Wroclaw, Poland

Co-authors: Damian Brzyski, Ulrike Schneider

The SLOPE estimator is defined as the minimizer of the penalized residual sum of squares where the penalty is the SLOPE norm (a generalization of the L1 norm). Because the objective function is not strictly convex the uniqueness of the minimizer is not obvious. We give a necessary and sufficient condition under which the uniqueness of the minimizer occurs. In addition, we show how a geometric condition involving the sign permutahedron gives insights about the accessible models for SLOPE estimator.

E2015: **Identifying aberrant EEG functional connectivity in schizophrenia using an ensemble of convolutional neural networks**

Presenter: **Chee Ming Ting**, King Abdullah University of Science and Technology, Saudi Arabia

Co-authors: Hernando Ombao

The aim is to leverage on altered patterns in brain functional connectivity as features for automatic discriminative analysis of neuropsychiatric patients. Deep learning methods have been introduced recently to fMRI functional network classification, however, existing architectures focused on a single type of connectivity measure. We propose a deep convolutional neural network (CNN) for classifying electroencephalogram (EEG)-derived brain connectome in schizophrenia (SZ). To capture complementary aspects of disrupted connectivity in SZ, we explore fusion of heterogeneous connectivity features consisting of time and frequency-domain metrics of effective connectivity based on vector autoregressive model and partial directed coherence, and complex network measures of network topology. We design a novel multi-domain connectome CNN (MDC-CNN) based on a parallel ensemble of 1D and 2D CNNs to integrate these features from various domains and dimensions using different fusion strategies. We also consider an extension to dynamic brain connectivity using the recurrent neural networks. Hierarchical latent representations learned by the multiple convolutional layers from EEG connectivity reveals apparent group differences between SZ and healthy controls (HC). Evaluated on resting-state EEG data, the proposed MDC-CNN by integrating information from diverse brain connectivity descriptors is able to accurately discriminate SZ from HC, outperforming support vector machines.

EO104 Room MAL B04 RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS Chair: Anne Ruiz-Gazen

E0508: **Regularized optimal transport of covariates and outcomes in data recoding**

Presenter: **Valerie Gares**, Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, Rennes, France

When databases are constructed from heterogeneous sources, it is not unusual that different encodings are used for the same outcome. In such case, it is necessary to recode the outcome variable before merging two databases. The method proposed for the recoding is an application of optimal transportation where we search for a bijective mapping between the distributions of such variable in two databases. Using that common covariates appear in the two databases, the objective is to minimize the expectation of a cost function reflecting a distance measure in the space of the covariates. The first form of the algorithm transports the distributions of categorical outcomes assuming that they are distributed equally in the two database. Then, we extend the scope of the model to treat all the situations where the covariates explain the outcomes similarly in the two databases. In particular, we do not require that the outcomes be distributed equally. For this, we propose a model where joint distributions of outcomes and covariates are transported. We also propose to enrich the model by relaxing the constraints on marginal distributions and adding an L1 regularization term. The performances of the models are evaluated in a simulation study, and they are applied to a real dataset.

E0561: **Asymptotically and computationally efficient tensorial JADE**

Presenter: **Joni Virta**, Aalto University, Finland

Co-authors: Niko Lietzen, Pauliina Ilmonen, Klaus Nordhausen

A novel method of tensorial independent component analysis is proposed based on TJADE and k-JADE, two recently proposed generalizations of the classical JADE algorithm. The new method achieves the consistency and the limiting distribution of TJADE under mild assumptions, and at the same time, it offers notable improvement in computational speed. The trade-off between computational speed and assumptions is controlled by a tuning parameter which has a natural interpretation as a maximal kurtosis multiplicity. Simulations and timing comparisons demonstrate the method's gain in speed and, moreover, the desired efficiency is obtained approximately also for finite samples. The method is applied successfully to large-scale video data, for which neither TJADE nor k-JADE is feasible.

E0933: **A flexible EM-like clustering algorithm for noisy data**

Presenter: **Violeta Roizman**, L2S (CentraleSupelec) - Instituto de Calculo (UBA), France

Co-authors: Frederic Pascal, Matthieu Jonckheere

A new robust clustering algorithm is designed that can deal efficiently with noise and outliers in diverse datasets. As an EM-like algorithm, it is based on both estimations of clusters centers and covariances, but also on a scale parameter per data-point. This allows the algorithm to accommodate for heavier/lighter tails distributions (in comparison to classical Gaussian distributions) and outliers without significantly loosing efficiency in classical scenarios. Convergence and accuracy of the algorithm are first analyzed by considering synthetic data. Then, we show that the proposed algorithm outperforms other classical unsupervised methods of the literature such as k-means, the EM algorithm and HDBSCAN when applied to real datasets as MNIST, NORB and 20newsgroups.

E1189: **Comparing prediction procedures for functional data in aeronautic**

Presenter: **Feriel Boulfani**, Mathematics institute of Toulouse, France

Co-authors: Anne Ruiz-Gazen, Xavier Gendre, Martina Salvignol

The oil temperature of an aircraft generator indicates the well functioning of the generator. By predicting the changes of the oil temperature, an abnormal behavior can be detected. For this reason, predicting the oil temperature behavior at a given time t is of particular interest. We propose to use as explanatory variables the observations of some auxiliary functions measured by the aircraft during a period preceding time t. The purpose is to consider a functional data analysis framework and compare it with the neural network, random forest and linear regression statistical procedures. To avoid overfitting, we apply the dropout technique to these procedures. This technique is used in neural network and consists of dropping randomly units in the network, which can be seen as a way of adding noise to the model. The dropout technique for linear regression model drops randomly some dimensions of the input while for random forest, it drops randomly some decision trees. The methods are compared using some real data set from the aeronautic sector.

E1418: **Cellwise robust and sparse regression with the shooting S**

Presenter: **Christophe Croux**, Edhec Business School, France

Co-authors: Ines Wilms, Lea Bottmer

Consider a high dimensional multiple regression model. The lasso is a popular estimator to reduce the dimensionality by imposing sparsity on the estimated regression parameters. As such, the lasso performs variable selection since it only keeps a few predictors and discards the remaining predictors by setting their respective parameter estimates to zero. The lasso is, however, not a robust estimator. Nevertheless, outliers, i.e. atypical observations, frequently occur in high-dimensional data sets. Therefore, we discuss a cellwise robust lasso estimator, the sparse shooting S. This estimator can deal with cellwise contamination, where many cells of the design matrix of the predictor variables may be outlying. Moreover, the sparse shooting S is computable in high-dimensional settings with more predictors than observations and it gives sparse parameter estimates. We compare its performance to several other sparse and/or robust regression estimators.

EO643 Room MAL B18 THE STATE-OF-THE-ART DEVELOPMENTS FOR NON-IGNORABLE MISSING DATA Chair: Shaun Seaman

E0776: **How not to estimate the nonignorable missingness mechanism**

Presenter: **Jiwei Zhao**, State University of New York at Buffalo, United States

Co-authors: Yanyuan Ma

The estimation problem is considered in a regression setting where the outcome variable is subject to nonignorable missingness and identifiability is ensured by the shadow variable approach. We propose a versatile estimation procedure where the modeling of the missingness mechanism is completely bypassed. We show that the proposed estimator is easy to implement, and we derive its asymptotic theory. We also investigate some alternative estimators under different scenarios. Comprehensive simulation studies are conducted to demonstrate the finite sample performance of the method. We apply the estimator to a children's mental health study to illustrate its usefulness.

E0934: **Methods of handling cohort data with death and non-ignorable dropout**

Presenter: **Lan Wen**, Harvard University, United States

Co-authors: Shaun Seaman

Three methods are proposed to model cohort data where repeated outcomes may be missing due to death and non-ignorable dropout. Examples arise in HIV studies where CD4 cell counts may be missing among those who are diagnosed with AIDS, or survey studies where some cognitive function outcomes are missing among the elderly. When missing data arise from death and dropout, one may want to distinguish between reasons for missingness to avoid making inferences about a cohort where no one can die. Instead, inferences based on those who are alive at any point in time might be more informative for health policy-makers. In order to obtain valid inference on those who are alive at each time point, we state the assumptions about the non-ignorable missingness process and we put forward: i) an inverse probability weighted method that upweights observed subjects to represent subjects who are still alive but are not observed; ii) an outcome regression method that replaces missing outcomes for subjects who are alive with their conditional mean outcomes given past observed data; and iii) an augmented inverse probability method that combines the previous two methods and is doubly robust against model misspecification. Through simulation, we compare the bias, efficiency and coverage probability of the three methods, and apply them to a cohort of elderly adults from the Health and Retirement Study.

E1224: **Identification, estimation, and semiparametric efficiency of nonignorable missing data**

Presenter: **Wang Miao**, Peking University, China

Identification and estimation with an outcome missing not at random will be discussed. We note that without extra assumptions, even fully parametric models are not identified. We then study identification strategies based on auxiliary variables such as an instrumental variable or a shadow variable. An instrumental variable impacts the missingness, but not the outcome of interest, and in contrast, a shadow variable is correlated with the outcome, but independent of the missingness after conditioning on the outcome. We describe general conditions for nonparametric identification of the full data law. We describe semiparametric estimation methods, and we characterize the semiparametric efficiency bound for the class of doubly robust regular and asymptotically linear estimators.

E1054: **Nonparametric regression with selectively missing covariates**

Presenter: **Christoph Breunig**, Humboldt-Universitat zu Berlin, Germany

Co-authors: Peter Haan

The problem of regressions with selectively observed covariates is considered in a nonparametric framework. The proposed approach relies on instrumental variables that explain variation in the latent covariates, but have no direct effect on selection. The regression function of interest is shown to be a weighted version of observed conditional expectation where the weighting function is a fraction of selection probabilities. Nonparametric identification of the fractional probability weight (FPW) function is achieved via a partial completeness assumption. We provide primitive functional form assumptions for partial completeness to hold. The identification result is constructive for the FPW series estimator. We derive the rate of convergence and also the pointwise asymptotic distribution. In both cases, the asymptotic performance of the FPW series estimator does not suffer from the inverse problem which derives from the nonparametric instrumental variable approach. In a Monte Carlo study and an empirical illustration, we analyze the finite sample properties of our estimator and we demonstrate the usefulness of our method in analyses based on survey data.

E0983: **Instrument, variable and model selection with nonignorable nonresponse**

Presenter: **Jun Shao**, University of Wisconsin - Madison, United States

With nonignorable nonresponse, an effective method to construct valid estimators of population parameters is to use a covariate vector, called instrument, that can be excluded from the nonresponse propensity, but contains still useful covariates, even when other covariates are conditioned. The existing work in this approach assumes such an instrument is given, which is frequently not the case in applications. We investigate how to search for an instrument from a given set of covariates. The method for estimation is the pseudo likelihood which assumes that the distribution of response given covariates is parametric and the propensity is nonparametric. Thus, in addition to the challenge of searching an instrument, we also need to do variable and model selection simultaneously. We propose a method for instrument, variable, and model selection and show that our method produces consistent instrument and model selection as the sample size tends to infinity, under some regularity conditions. Empirical results including two simulation studies and two real examples are present to show that the proposed method works well.

EO162 Room MAL B20 RECENT ADVANCES IN NETWORK DATA ANALYSIS Chair: Yuan Zhang

E1500: **Coauthorship and citation networks for statisticians**

Presenter: **Jiashun Jin**, Carnegie Mellon University, United States

A data set has been collected for the networks of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of papers published in representative journals, for 36 journals in statistics and related fields, spanning about 35 years. The data set provides a fertile ground for research in social networks, text mining, and knowledge discovery, and motivates an array of interesting problem in statistics and machine learning. We provide an expository overview on this data set, and discuss several problems including overall productivity of statisticians, statistical journal ranking, citation patterns, co-authorship network communities, co-authorship network mixed-memberships, dynamic networks, and topic estimation.

E0591: **Estimation and clustering in popularity adjusted stochastic block model**

Presenter: **Marianna Pensky**, University of Central Florida, United States

Co-authors: Majid Noroozi, Ramchandra Rimal

Stochastic networks in general and stochastic block models in particular attracted a lot of attention in the last decade. The Popularity Adjusted Block model (PABM) is considered, which generalizes the Stochastic Block model and the Degree Corrected Block Model by allowing more flexibility for block probabilities. We argue that the main appeal of the PABM is its less rigid spectral structure which makes the PABM an attractive choice for modeling networks that appear in biological sciences.

E1378: **Optimal adaptivity of signed-polygon statistics for network testing**

Presenter: **Tracy Ke**, Harvard University, United States

Given a symmetric social network, the focus is on testing whether it has only one community or multiple communities. The desired tests should (a) accommodate severe degree heterogeneity, (b) accommodate mixed-memberships, (c) have a tractable null distribution, and (d) adapt automatically to different levels of sparsity, and achieve the optimal phase diagram. How to find such a test is a challenging problem. We propose the Signed Polygon as a class of new tests. Fixing $m > 2$, for each m-gon in the network, we define a score using the centered adjacency matrix. The sum of such scores is then the m-th order Signed Polygon statistic. The Signed Triangle (SgnT) and the Signed Quadrilateral (SgnQ) are special examples of the Signed Polygon. We show that both the SgnT and SgnQ tests satisfy (a)-(d), and especially, they work well for both very sparse and less sparse networks. The analysis of the SgnT and SgnQ tests is delicate and tedious, and the main reason is that we need a unified proof that covers a wide range of sparsity levels and a wide range of degree heterogeneity. For lower bound theory, we use a phase transition framework, which includes the standard minimax argument, but is more informative.

E0500: **Network structure inference from grouped data**

Presenter: **Yunpeng Zhao**, Arizona State University, United States

Statistical network analysis typically deals with inference concerning various parameters of an observed network. In several applications, especially those from social sciences, behavioral information concerning groups of subjects are observed. In such data sets, even though a network structure is present it is not typically observed. These are referred to as implicit networks. We describe a model-based framework to uncover the implicit network structure and address related inferential questions. We also describe extensions of the methodology to time series of grouped observations.

E1355: **From graph structure to network function: Using spectral graph theory to predict and control dynamics**

Presenter: **Rosemary Braun**, Northwestern University, United States

Co-authors: Phan Nguyen

The structure of networked system governs the dynamics of processes taking place on the graph (such as the flow of current, information, etc.) Spectral decomposition of the graph Laplacian provides a means to summarize the network structure and make predictions about those dynamics. Spectral graph theory (SGT) has been used extensively to analyze networked systems, cluster data, and perform dimension reduction. However, much the underlying theory was developed in the context of unsigned, undirected graphs; in contrast, real networks are often directed/asymmetric, and have both positive ("activating") and negative ("inhibitory") valences. We will discuss our recent efforts to extend spectral graph theory to directed, signed networks. We will also describe a new SGT-based approach for predicting dynamics on the graph from static observations, and for identifying network elements that can be targeted to control those dynamics. Finally, we will demonstrate how these methods can be exploited to restore the spectral (and hence the dynamical) properties of a "damaged" network, even when the original graph topology cannot be recovered.

EO146 Room MAL B35 MULTIVARIATE SURVIVAL MODELS Chair: Takeshi Emura

E0975: **A Gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker**

Presenter: **Krithika Suresh**, University of Colorado, United States

Co-authors: Jeremy Taylor, Alexander Tsodikov

Dynamic prediction uses patient information collected during follow-up to produce individualized survival predictions at given time points beyond treatment or diagnosis. This allows clinicians to obtain updated predictions of a patient's prognosis that can be used in making personalized treatment decisions. Two commonly used approaches for dynamic prediction are landmarking and joint modeling. Landmarking does not constitute a comprehensive probability model, and joint modeling often requires strong distributional assumptions and computationally intensive methods for estimation. We introduce an alternative approximate approach for dynamic prediction that aims to overcome the limitations of both methods while achieving good predictive performance. We separately specify the marker and failure time distributions conditional on surviving up to a prediction time of interest and use standard variable selection and goodness-of-techniques to identify the best-fitting models. Taking advantage of its analytic tractability and easy two-stage estimation, we use a Gaussian copula to link the marginal distributions smoothly at each prediction time using an association function. With simulation studies, we examine the proposed method's performance and identify situations where it is preferable over existing approaches. We illustrate its advantages for dynamic prediction in an application to health data.

E1366: **Nonparametric inference for copulas of dependence under length-biased sampling and informative censoring**

Presenter: **Taoufik Bouezmarni**, Universite de Sherbrooke, Canada

Co-authors: Yassir Rabhi

Length-biased sampling is common in cross-sectional surveys and prevalent- cohort studies, and is well known to induce bias on the samples variables. The truncation mechanism in such sampling tends to over-select large values and under-select small values of some variables (e.g. lifetime). We consider copulas for modeling the dependence when the collected data are length-biased and subject to informative censoring. For such data, where large values of some variables are more frequent than small ones, modeling the dependence structure without correction leads to biased results. We address nonparametric estimation of the bivariate distribution, copula function and its density, and Kendall and Spearman measures for right-censored length-biased data. The proposed estimator for the bivariate cdf is a Hadamard-differentiable functional of two MLEs (Kaplan-Meier and empirical distributions). Based on this estimator, we devise two estimators for copula function and a local-polynomials estimator for copula density, that accounts for boundary bias. Also, we introduce estimators for Kendall and Spearman measures. The limiting processes of the estimators are established by deriving i.i.d. representations. As by-product, we establish the oscillation behavior of the bivariate cdf estimator.

E1155: **Estimating dementia incidence in an illness-death model using health-claims data**

Presenter: **Achim Doerre**, University of Rostock, Germany

The motivation comes from the study of dementia in humans, which usually does not occur before reaching age 60. We use an illness-death model to describe the health process over time, where individuals migrate from the initial healthy (non-diseased) state to either the death or disease state. While the death state is absorbing, migration to the disease state is eventually followed by migration to the death state. This model contains three migration rates that are unknown in general, among which the incidence rate is of primary interest. When incidence rates of chronic diseases in humans are studied, cohort and period effects are often of interest. Unfortunately, simple random samples are usually not obtainable for age-related diseases. As an alternative, health-claims data offer large-scale contemporary information on the health status of individuals, and may be regarded as observational data. Under this sampling scheme, incidence times are both left and right-censored. Furthermore, truncation occurs because only those individuals are sampled which are alive when the data collection begins, leading to individual-specific truncation times. We derive consistent and asymptotically normal Maximum Likelihood estimators using an inverse probability weighted likelihood function. In order to account for

different cohorts possibly sharing certain migration rates, we describe a simple model selection procedure. A large health-claims dataset is used to illustrate the method.

E0636: **The identifiability of the copula competing risks model under exclusion restrictions**
Presenter: **Ralf Wilke**, Copenhagen Business School, Denmark
Co-authors: Simon Lo

Competing risks duration models are routinely applied in many disciplines such as biostatistics, mechanical engineering, economics and other social sciences. A convenient and general way to describe risk dependencies is to model the joint distribution by a copula that links the marginal distributions of latent competing durations. While the non-identifiability of the competing risks model complicates informative empirical analysis, a series of contributions has obtained identification results under different sets of restrictions. We develop a new identification result under exclusion restrictions. We show that the existence of exclusion restrictions ensures model identification under mild additional restrictions. While exclusion restrictions are common in econometric models, previous identification results for competing risks models do not rely on them.

E0954: **Kendalls tau for survival endpoints in meta-analysis: A general definition and a conditional copula approach**
Presenter: **Takeshi Emura**, National Central University, Taiwan
Co-authors: Virginie Rondeau, Sofeu Casimir

Measuring dependence between survival endpoints is an essential process to understand the effect of treatments in clinical trials. In particular, an individual-patient data (IPD) meta-analysis for validating a surrogate endpoint requires the estimation of individual-level dependence between the surrogate endpoint and the true endpoint. Copula models and frailty models have been suggested as promising tools to compute Kendall's tau between the surrogate endpoint and the true endpoint. However, these model-based approaches for computing Kendall's tau seem to impose a simplifying assumption: That is, Kendall's tau does not depend on treatment arms. We propose a general definition of Kendall's tau that allows different values of tau across different treatment arms. We argue that the simplifying assumption in the existing models is questionable for measuring dependence between progression-free survival and overall survival. Motivated by these findings, we propose a frailty-conditional copula model for the IDP meta-analysis with two survival endpoints. We examine the performance of the proposed methods via simulations. The proposed method is implemented in an R package joint.Cox available in CRAN.

EO158 Room MAL B36 RECENT DEVELOPMENTS ON COMPLEX DATA ANALYSIS Chair: Man Wang

E0734: **MOSUM-based test and estimation method for multiple changes in panel data**
Presenter: **Man Wang**, Donghua University, China

Common change point detection is a vibrant topic in panel data analysis, however, most existing tests are based on the cumulative sum (CUSUM) method and suffer power loss under certain multiple change points setting. To solve this problem, a moving sum (MOSUM) based test is proposed to detect the common breaks happened in panel data with cross sectional dependence. Under mild conditions, it is shown that the proposed test statistic converges to an extreme distribution of a Gaussian process under the null hypothesis and diverges to infinity under the alternative hypothesis. Numerical studies show that the proposed method outperforms existing CUSUM-based procedures under multiple change points setting, as well as the single change point case with small sample size. We also give an estimation method based on the proposed test and establish its consistency. Application to the US state-level personal income data is also demonstrated.

E0746: **Sample size determination guidelines for latent mediation models: A Monte Carlo study**
Presenter: **Rongqian Sun**, The Chinese University of Hong Kong, China
Co-authors: Junhao Pan, Jingheng Cai

Mediation analysis has been extensively applied to psychology and many other disciplines to explore the intrinsic influencing mechanism underlying phenomena or behaviors that researchers are interested in. An important issue researchers frequently encounter is: How many samples are needed to ensure sufficient power of testing the mediation effect?. Several empirical rules have been proposed for simple mediation models with manifest variables. Despite latent mediation analysis covers the measurement of unobservable constructs and is therefore much more widely used in practice, there is a paucity of study on the sample size required for testing latent mediation effects. The main objective is to provide empirically supported sample size guidelines for several representative latent mediation models commonly used in applied researches, which is helpful for experimental design and research funding application in psychology and so on. Monte Carlo simulation experiments are conducted to investigate minimum sample size required for power no less than 0.8 in the test of latent mediation effect under 144 modeling conditions constituted by influencing factors including the number of items per factor, effect size, and estimation methods. Besides, the comparison between frequentist and Bayesian approach is extended to the context of latent mediation analysis

E1029: **Order selection in mixed hidden Markov model**
Presenter: **Yiqi Lin**, The Chinese University of Hong Kong, Hong Kong

In the recent decades, Hidden Markov model (HMM) is widely used in many research fields. However, traditional HMMs frequently assume that the number of hidden states (order of HMM) is a constant and should be specified prior to analysis. This assumption is apparently unrealistic and too restrictive in many applications. We consider HMMs by allowing the number of hidden states to be unknown and determined by the data. We propose a novel likelihood-based penalized method, along with an efficient Monte Carlo expectation conditional maximization (MCECM) algorithm, to simultaneously perform order selection and parameter estimation in the context of HMMs. Simulation studies are conducted to evaluate the performance of the proposed method. An application of the proposed methodology to a real-life study is presented.

E1796: **Bayesian single-cell transcriptome analysis and related MCMC convergence diagnostics**
Presenter: **Sheng Lian**, The Chinese University of Hong Kong, China

Single-cell RNA sequencing has become widely used in recent years. The ability to measure the gene expression at a cellular resolution allows us to study new biological questions related to cell-specific changes. However, there exists a large amount of technical noise that may hinder downstream analysis. Dropout events happen when a gene is not detected owing to a failure to capture or amplify. Based on the assumption that the detection rate for each gene in every single cell depends on the level of expression, we propose a hierarchical model with the non-ignorable missing-data mechanism to model the dropout events. Bayesian inference based on Markov Chain Monte Carlo (MCMC) algorithms is performed. Also, trace plot for the posterior of the joint parameter is used for diagnosing convergence problems and some unexpected behaviors are observed. Then, we investigate the reasons in hierarchical modeling framework and provide guidance in MCMC convergence diagnostics. Finally, we move into single-cell cancer genomic studies and work on the method to identify differentially expressed genes between normal and diseased conditions. We conduct statistical analysis that accounts for the dropout events and illustrate how this step help us better understand disease heterogeneity at the single-cell level.

E1805: **Second-order numerical Fourier methods for option pricing**
Presenter: **Jinhui Han**, The Chinese University of Hong Kong, China

A Fourier method will be applied to numerically simulate the option prices of various types, based on the second-order stochastic Taylor discretization of the underlying SDE dynamics. Particularly, we focus on high-dimensional SDEs, where the stochastic Levy area will be involved in the second-order expansion. The theoretical one-step conditional characteristic function is derived. To better simulate the characteristic function, a neural network is adopted and served as an interpolation function. The method differs from traditional Monte-Carlo simulations where each step

requires an independent sampling process. Instead, the efficient Fourier method enables us to sample the characteristic function values on the grids for only once and it can be settled as an online library which is fully determined by the coefficients of the corresponding SDE. In this way, a higher weak convergence order is achieved and not much additional computational efforts will be required. The error analysis is conducted in detail as well. Finally, numerical experiments for computing different options prices are provided, including European options as well as path-dependent ones such as Bermuda options.

EO070 Room MAL G13 EMPIRICAL PROCESSES AND NONPARAMETRIC METHODS Chair: Eric Beutner

E1174: **Differentiability of supremum-type functionals with applications**
Presenter: **Javier Carcamo**, Universidad Autonoma de Madrid, Spain
Co-authors: Antonio Cuevas, Luis-Alberto Rodriguez

The purpose is to show that various functionals related to the supremum of a real function defined on an arbitrary set or a measure space are Hadamard directionally differentiable. We specifically consider the supremum norm, the supremum, the infimum, and the amplitude of a function. The (usually non-linear) derivatives of these maps adopt simple expressions under suitable assumptions on the underlying space. As an application, we improve and extend to the multidimensional case previous results regarding the limiting distributions of Kolmogorov-Smirnov type statistics under the alternative hypothesis. Similar results are obtained for analogous statistics associated with copulas. We additionally solve an open problem about the Berk-Jones statistic. Finally, the asymptotic distribution of maximum mean discrepancies over Donsker classes of functions is derived.

E1620: **Randomization empirical processes and deduced hypothesis tests**
Presenter: **Dennis Dobler**, Vrije Universiteit Amsterdam, Netherlands

Paralleling results for the bootstrap and random permutation, a broadly applicable conditional Donsker theorem is developed for empirical processes which are based on randomized observations. Random elements of an algebraic group are applied to the raw data and a statistic is applied to the randomized data. After a suitable studentization of the statistic, asymptotically exact hypothesis tests are deduced. They are even finitely exact under group-invariant hypotheses. The methodology is exemplified with a right-censored paired data problem and the question: which entry has the better survival chances? To this end, a suitable Wilcoxon-type test will be developed.

E1424: **Likelihood ratio tests and confidence intervals based on the shape constraint of concavity**
Presenter: **Charles Doss**, University of Minnesota, United States
Co-authors: Jon A Wellner

Estimation and inference for a log-concave density and for a concave regression function is considered. These problems have some similarities because they both rely on an underlying shape constraint of concavity. Forming confidence intervals or hypothesis tests in nonparametric settings is often challenging. We propose using likelihood ratio statistics to form hypothesis tests (which can be inverted to form confidence intervals). We consider tests or intervals for the location of the mode of the log-concave density function and for the value of the concave regression function. The statistics we propose are tuning parameter free, a rarity in nonparametric settings. We demonstrate that the likelihood ratio statistics are asymptotically pivotal (satisfy the so-called Wilks phenomenon). Thus, they have universal critical values not depending on any unknown parameters, allowing the tests or intervals to be computed in practice.

E1462: **Empirical process for merged data from multiple overlapping sources**
Presenter: **Takumi Saegusa**, University of Maryland, United States

Empirical process theory is studied for merged data from multiple overlapping data sources. The setting we consider is characterized by (1) heterogeneity of multiple data sets, (2) unidentified duplication across samples, (3) dependence due to finite population sampling. The resultant sample is a biased and dependent sample with duplication. The standard empirical process theory often assumes an independent and identically distributed sample, and hence most results do not hold in this setting. We develop the uniform law of large numbers and uniform central limit theorem for data integration. We apply these empirical process results to general theorems for consistency, rates of convergence and asymptotic normality of infinite dimensional M-estimators. The results are illustrated with simulation studies and a real data example using the Cox proportional hazards model.

E1290: **Donsker results for the smoothed empirical process**
Presenter: **Henryk Zaehle**, Saarland University, Germany
Co-authors: Eric Beutner

New results are presented on convergence in distribution of the smoothed empirical process on the real line for fairly large index sets of functions under weak assumptions. The results allow for a flexible choice of the bandwidth and cover short-range dependence of the underlying data. The results continue to hold under long-range dependence when the central rate is replaced by a suitable non-central rate.

EO681 Room MAL G14 RISK, VARIABILITY AND HEAVY TAILS Chair: Anna Panorska

E0748: **Weibull distribution in insurance risk models**
Presenter: **Corina Constantinescu**, University of Liverpool, United Kingdom

A few instances of using Weibull distributions in modelling risk events (frequency or intensity) in mathematical insurance models will be discussed.

E0770: **Heavy-tailed stochastic models with time-dependent coefficients: Applications to financial time series**
Presenter: **Agnieszka Wylomanska**, Wroclaw University of Science and Technology, Poland

To properly manage market risk, industrial companies use tools based on value-at-risk, which requires proper modeling of future risk factors dynamics. One of the major challenges faced by this technique applied to modeling financial time series is the choice of an appropriate model for the simulation of the future paths. In order to fit the data, it should reflect its properties including heavier than Gaussian tail distributions, stabilizing volatility and mean reversion in long term horizon. We propose to apply the extension of the classical stochastic model with fixed coefficients for the description of currency exchange rates data and the metals prices. This model was introduced for describing the evolution of the short interest rate and could be considered as the natural extension of the classical Ornstein-Uhlenbeck process, where the coefficients are constant. The standard version of the model was based on the Brownian motion (BM). However, it can be easily extended to any class of distributions. Since the financial data of interest exhibit non-Gaussian behavior, we modify the model to use skewed generalized Students t-distribution. We demonstrate the estimation techniques and present the real-life applications to financial time series.

E1373: **Non-Gaussian harmonizable processes as sum of harmonics with random frequencies**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Krzysztof Podgorski

A class of strictly stationary non-Gaussian stochastic processes is discussed that allows to independently specify the spectral and marginal first-order distribution of the process. The proposed method models the sinusoidal component frequencies as random variables with a distribution specified by the spectral distribution. This is a key departure from the classical representation of a stationary process by the spectral theorem. While it is known that this class belongs to the non-ergodic Harmonizable Processes (HP), we identify the field of invariant sets, and thus, also the limiting behavior

of the time averages of functionals of such a process. The ergodic properties of the derived models are detailed for the class of such processes with G-type marginal distributions.

E1532: **A computational approach for the estimation of discrete Pareto parameters**
Presenter: **Tomasz Kozubowski**, University of Nevada Reno, United States
Co-authors: Charles Amponsah
A discrete Pareto distribution is a probability model with a power-law tail, which provides a convenient alternative to the well-known Zipf distribution. We propose a computational approach to parameter estimation connected with this model viz. the expectation-maximization (EM) algorithm. The approach is illustrated by a simulation study and a real data example from finance.

E1572: **Spatio-temporal dependence measures for two-dimensional VAR(1) models with alpha-stable innovations**
Presenter: **Marek Teuerle**, Wroclaw University of Science and Technology, Poland
Many real phenomena exhibit non-Gaussian behavior. The non-Gaussianity is manifested by an impulsive behavior of the real data that can be found in both one-dimensional and multi-dimensional cases. Especially, the multi-dimensional datasets with non-Gaussian behavior pose substantial analysis challenges to scientists and statisticians. We analyze the bidimensional vector autoregressive (VAR) model based on general bidimensional alpha-stable distribution. This time series can be applied in modeling bidimensional data with impulsive behavior. We focus on the description of the spatio-temporal dependence for analyzed bidimensional time series which in the considered case cannot be expressed in the language of the classical cross-covariance or cross-correlation function. We propose a new cross measure based on the alternative measure of dependence adequate for infinite variance processes, namely cross-covariation. We provide an extension of the authors' previous work where the cross-codifference was considered as the spatio-temporal measure of the components of VAR model based on sub-Gaussian distribution. We demonstrate that cross-codifference and cross-covariation can give different information about the relationships between components of bidimensional VAR models.

E1503: **Probabilistic properties of detrended fluctuation analysis for Gaussian processes**
Presenter: **Grzegorz Sikora**, Wroclaw University of Science and Technology, Poland
The detrended fluctuation analysis is one of the most widely used tools for the detection of long-range dependence in time series. Although DFA has found many interesting applications and has been shown as one of the best performing detrending methods, its probabilistic foundations are still unclear. We study the probabilistic properties of DFA for Gaussian processes. The main attention is paid to the distribution of the squared error sum of the detrended process. This allows us to find the expected value and the variance of the squared fluctuation function of DFA for a Gaussian process of a general form. The results can serve as a starting point for analyzing the statistical properties of the DFA-based estimators for the fluctuation function and long-memory parameters. The obtained theoretical results are supported by numerical simulations.

EO328 Room MAL G15 BAYESIAN METHODS IN MEDICAL STATISTICS Chair: Paulo Canas Rodrigues

E1100: **Distributional ROC surface regression**
Presenter: **Vanda Inacio**, University of Edinburgh, United Kingdom
Co-authors: Nadja Klein
Accurate diagnosis of disease is of great importance in clinical practice and medical research. The receiver operating characteristic (ROC) surface is a popular tool for evaluating the discriminatory ability of continuous diagnostic test outcomes when there exist three ordered disease classes (e.g., no disease, mild disease, advanced disease). Incorporating covariates in the analysis can potentially enhance information gathered from the diagnostic test, as its discriminatory ability may depend on these. We propose a Bayesian distributional regression approach for covariate-specific ROC surface estimation. In the model specification, the covariate-specific ROC surfaces are indirectly modelled using probabilistic distributional models capturing location, scale, shape, and other aspects of the diagnostic test distribution in each of the three groups, where covariate effects are modelled through penalised splines. Multiple simulation studies demonstrate the ability of the model to successfully recover the true covariate-specific ROC surface and the corresponding covariate-specific VUS in a variety of complex scenarios. The methods are motivated by and applied to a prostate cancer study where the main goal is to assess if and how the accuracy of a new diagnostic test, the prostate health index density, changes with age.

E0339: **Biomarker accuracy determination via an affinity-based measure**
Presenter: **Bradley Barney**, University of Utah, United States
Co-authors: Miguel de Carvalho, Garritt Page
The area under the receiver operating characteristic curve (AUC) and Youden index are popular metrics for quantifying the ability of a biomarker to predict the presence/absence of a disease. These metrics can perform poorly when the biomarker distribution with/without the condition does not stochastically dominate the other. We propose using a measure based on Hellinger's affinity which overcomes this potential pitfall, providing a global summary of the similarity between biomarker densities from individuals with versus without a disease. We also illustrate the application of nonparametric Bayesian methods to flexibly estimate a covariate-dependent version of the affinity-based measure. We investigate the performance of our suggested metric via simulation. We also apply the methodology to a prostate cancer study to compare the potential diagnostic ability of two prostate-specific antigen (PSA)-derived biomarkers.

E0767: **Parallelising Bayesian biostatistical analyses using Yin-Yang sampling**
Presenter: **Alexandra Posekany**, University of Technology Vienna, Austria
Co-authors: Sylvia Fruhwirth-Schnatter
For decades, Bayesian methods have been more widely applied in biostatical and medical analyses. A part of that success is their ability to integrate new observations with previous knowledge from other studies or experts. However, the drawback of good handling of small samples is that Bayesian models for large studies analysing 100000s of patients from medical registries suffer from the required computational intensity w.r.t. memory usage and need for parallelisation. To get a step closer to performing such large calculations, we developed a methodology which is splits the data into smaller subsets, performing Bayesian inference independently on each and finally merges the separate results into a joint merged result which is comparable to the inference of the complete data set. The method was named Yin-Yang sampling, because it focusses on merging information from two different sources-the yin and the yang sample-by correcting multiple use of the prior information. By appling yin-yang sampling steps, we recover the full samples posterior from any given number of subsamples posteriors with some restrictions. For demonstration, an inference with logistic regression on a data set from the Austrian Stroke registry containing over 100000 patients is shown. This provides a scenario where the full samples inference is impossible on a desktop computer due to lack of memory, while subsample computation plus the Yin-Yang merging algorithm require only minutes for computation.

E0358: **Bayesian analysis of survival data with missing censoring indicators**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
In some large clinical studies, it may be impractical to perform the physical examination to every subject at his/her last monitoring time in order to diagnose the occurrence of the event of interest. This gives rise to survival data with missing censoring indicators where the probability of missing may depend on time of last monitoring and some covariates. We present a fully Bayesian semi-parametric method for such survival data to estimate regression parameters of Coxs proportional hazards model. Theoretical investigation and simulation studies show that the proposed method

performs better than competing methods. We apply this method to data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study.

E1658: **Bayesian nonparametric clustering of continuous-time hidden Markov models for health trajectories**
Presenter: **Yu Luo**, McGill University, Canada
Co-authors: David Stephens, David Buckeridge

Clustering procedures are developed for healthcare trajectories based on a continuous-time hidden Markov model and a generalized linear observation model. Specifically, we carry out Bayesian nonparametric inference for a Dirichlet process mixture model, and utilize restricted Gibbs sampling split-merge proposals to achieve inference using Markov chain Monte Carlo. In the analysis on a large Canadian cohort of subjects suffering from chronic obstructive pulmonary disease, a three-cluster model is chosen. The inferred Markov transition rate matrix in each cluster suggests that each cluster has its own transition characteristics and observation process, and that patients are more likely to stay in more severe disease states.

EO126 Room MAL G16 COPULAS AND DEPENDENCE MODELLING Chair: Piotr Jaworski

E0423: **Some transformations of bivariate independence copula**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania
Co-authors: Gediminas Bagdonas

Necessary and sufficient conditions on f for the function $H_f(C)(x,y) = C(x,y)f(1-x-y+C(x,y))$, $x,y \in [0,1]$ to be a bivariate copula for *any* bivariate copula C are known. If, on the other hand, $C(x,y) = \Pi(x,y) = xy$ is fixed, then some of those conditions become no longer necessary, presenting an interesting problem. Sufficient conditions, which unify several known examples, will be provided, and a discussion on their necessity, as well as on several interesting properties of C_f when it is indeed a copula, will be given.

E0682: **On Archimax copulas and dependence measures**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

The dependence measures, like for example Kendall tau, Spearman rho, Blomquist beta or tail dependence coefficients, are the main numerical characterization of Bivariate Archimax Copulas, a broad class of copulas containing Archimedean, Extreme Value and Conic Copulas. Such copulas are determined by a generator of an Archimedean copula and a function on the unit segment, called a Pickands dependence function, which is convex and comprised between two bounds. We identify the smallest possible compact sets containing the graphs of all Pickands dependence functions whose corresponding Archimax Copulas with given generator have fixed values of a dependence measure. We also provide the bounds for such sets of copulas.

E0695: **A Markov product for tail dependence functions**
Presenter: **Christopher Strothmann**, TU Dortmund University, Germany

A d-fold product structure * on the set of all bivariate tail dependence functions is introduced which is akin to the Markov product of 2-copulas. Several algebraic properties of * are investigated and its relationship to the tail behaviour of the Markov product of 2-copulas is examined. Finally, dependence reducing properties with regard to the extremal ordering are derived.

E0829: **Omnibus test for covariate effects in conditional copula models**
Presenter: **Irene Gijbels**, Katholieke Universiteit Leuven, Belgium
Co-authors: Marek Omelka, Noel Veraverbeke

Conditional copulas describe the conditional dependence and the influence that covariates have on the dependence structure between two (or more) variables. Of interest is to test the null hypothesis that the covariates have a specific effect (linear or quadratic effects, partial effects, ...). A special case is the testing problem that the covariates do not effect the dependence structure. We discuss an omnibus test for testing the null hypothesis of a specified effect of the covariates. The test statistic is designed for having power against many alternatives, and can be used to test for a variety of covariate effects (no effects, linear effects, partial effects, ...). In this semiparametric framework the marginal distribution functions are estimated using nonparametric kernel techniques and the parametric dependence model is estimated using maximum likelihood estimation. The asymptotic distribution of the test statistic under the null hypothesis is established. The finite-sample performance of the test is investigated via a simulation study. A real data analysis illustrates the practical use of the test.

E1477: **Extremal problem in a class of copulas**
Presenter: **Fabrizio Durante**, University of Salento, Italy
Co-authors: Juan Fernandez Sanchez, Manuel Ubeda Flores

Several optimization problems arising in the study of copula-based stochastic models have motivated the study of the extreme points (in the Krein-Milman sense) in the class of copulas. We focus on the class of extreme biconic copulas and provide their characterization. Moreover, we show how this characterization may help in shedding light upon new aspects of biconic copulas, including their link with extreme-value copulas.

EO769 Room CLO 101 LONGITUDINAL DATA ANALYSIS Chair: Sanjoy Sinha

E0207: **Analysis of survival and acquired over time treatment toxicities in elderly HNSCC patients: The SEER-Medicare database**
Presenter: **Olga Goloubeva**, University of Maryland School of Medicine, United States

The objective was to evaluate impact of primary treatment on survival and QOL outcomes of elderly patients with head and neck cancer (HNSCC). The SEER-Medicare linked database for those diagnosed with HNSCC during 1992-2011 was utilized. We identified all eligible patients 66 years and older with locally advanced cancers of the oropharyngeal, oral cavity and laryngeal SCC. The primary treatment defined as therapy received over first 180 days since cancer diagnosis. The probability of receiving single modality vs multimodality treatment was assessed by logistic regression approach. The multivariable Cox regression model stratified by cancer stage and site was used to estimate impact of any single modality vs multimodality on overall survival and toxicities time trend when adjusted for age and comorbidity index (up to 47 conditions) at time of diagnosis. The confounding bias in associating treatment with outcomes was accounted for by using propensity scoring. The study population comprised 5879 beneficiaries. Patients who were older and had higher CCI scores were less likely to receive more intense multimodality treatment across the three HNSCC sites. Elderly locally advanced HNSCC patients, particularly in the oropharynx and larynx, who receive less intense therapy have decreased survival. While patient selection is likely a co-factor in explaining these results, our findings are consistent with an age-independent benefit of getting multi-modality therapy.

E0219: **Modeling longitudinal data with measurement error in covariates**
Presenter: **Mahmoud Torabi**, University of Manitoba, Canada
Co-authors: Erfanul Hoque

Longitudinal data occur frequently in practice such as medical studies and life sciences. Generalized linear mixed models (GLMMs) are commonly used to analyze such data. It is typically assumed that the random effects covariance matrix is constant among subjects in these models. In many situations, however, the correlation structure may differ among subjects and ignoring this heterogeneity can lead to biases in model parameters estimate. Covariates measured with an error also happen frequently in the longitudinal data set-up (eg, blood pressure and cholesterol level). Ignoring this issue in the data may produce bias in model parameters estimate and lead to wrong conclusions. We propose an approach to properly

model the random effects covariance matrix based on covariates in the class of GLMMs, where we also have covariates measured with error. The resulting parameters from the decomposition of random effects covariance matrix have a sensible interpretation and can be easily modeled without the concern of positive definiteness of the resulting estimator. Performance of the proposed approach is evaluated through simulation studies and also by a real data application.

E0372: **Inference with joint models under misspecified random effects distributions**
Presenter: **Abdus Sattar**, Case Western Reserve University, United States
Co-authors: Sanjoy Sinha

Joint models are commonly used in clinical studies for analyzing survival data with time-dependent covariates or biomarkers. It is often assumed that the latent processes that are used to describe the association between longitudinal and survival outcomes follow a multivariate normal distribution. While a joint likelihood analysis may provide valid inferences under correctly specified latent processes or random effects distributions, the maximum likelihood estimators can be biased under misspecified random effects and hence may provide invalid likelihood inferences. We explore the empirical properties of the maximum likelihood estimators in joint models under various types of random effects distributions, and propose a robust and efficient skew-normal distribution to address uncertainties in the latent random effects distributions. An extensive Monte Carlo study indicates that the proposed method provides consistent and efficient estimators of the joint model parameters under various types of model misspecifications. We also present an application of the proposed method using a large clinical dataset obtained from the genetic and inflammatory markers of sepsis (GenIMS) study.

E0387: **Semiparametric methods for incomplete binary longitudinal data with dropouts**
Presenter: **Sanjoy Sinha**, Carleton University, Canada

Some semiparametric methods are discussed for joint estimation of the regression parameters and association parameters in binary longitudinal models when the marginal mean response function is partially linear. We propose a spline regression method in the framework of the weighted generalized estimating equations (GEEs) for the simultaneous estimation of the unknown nonlinear function, regression and association parameters under the assumption of a missing at random (MAR) dropout mechanism. Empirical properties of the proposed estimators are investigated using Monte Carlo simulations. An application is also provided using actual longitudinal data from a clinical study, where the data show strong evidence of a nonlinear trend in the mean response function.

E0805: **Statistical challenges in integrating survey and administrative data in the Canadian census**
Presenter: **Karelyn Davis**, Statistics Canada, Canada

Many national statistical offices are conducting research to better utilize administrative records, defined as data collected by government agencies or commercial businesses as part of administering a program or service. Administrative records offer the possibility to mitigate the lack of sustainability of the traditional survey enumeration approach, to offset declining response rates and to potentially reduce survey costs. Often, administrative records are longitudinal in nature and some European countries have incorporated such records into their national census, either wholly or partially, through the use of statistical or administrative registers. In these countries, estimation is subsequently based on information compiled in the registers as opposed to traditional survey enumeration. In Canada, a combined census approach is under research, whereby administrative data and traditional data collection are used jointly to enumerate populations under study. Such integration provides interesting statistical challenges, particularly with respect to record linkage methodology, estimation and hypothesis testing. Ongoing research into these challenges will be discussed, particularly how to form households from longitudinal individual administrative data in the Canadian context.

EO853 Room CLO 102 DATA SCIENCE: REGULARIZATION AND VARIABLE SELECTION Chair: Jia Liu

E0639: **Regularized area-level models for robust small area estimation under measurement errors**
Presenter: **Jan Pablo Burgard**, Trier University, Germany
Co-authors: Dennis Kreber, Joscha Krause

An approach is presented to model-based small area estimation under covariate measurement errors. Using a min-max approach, we prove that regularized regression coefficient estimation is equivalent to robust optimization under additive noise. Applying this equivalence, the Fay-Herriot model is extended by the l1-norm, the squared l2-norm, and elastic net regularizations as robustification against design matrix perturbations. This allows for reliable area-statistic estimates without distributive information about the measurement errors. A best predictor and a Jackknife estimator of the mean squared error are presented. The methodology is evaluated in a simulation study under multiple measurement error scenarios to support the theoretical findings. A comparison to other robust small area approaches is conducted. An empirical application to poverty mapping in the US is provided. Estimated economic figures from the US Census Bureau and crime records from the Uniform Crime Reporting Program are used to model the number of citizens below the federal poverty threshold.

E0642: **Regularized multi-level models for small area estimation using both unit- and area-level data**
Presenter: **Joscha Krause**, Trier University, Germany
Co-authors: Jan Pablo Burgard, Ralf Muennich

The joint usage of unit- and area-level data for model-based small area estimation is investigated. The combination of levels within a single model encloses a variety of methodological problems. Firstly, it implies a critical decrease in degrees of freedom due to more model parameters that need to be estimated. This may destabilize model predictions in the presence of small samples. Secondly, unit- and area- level data has different distributional characteristics in terms of dispersion patterns and correlation structure. Thirdly, unit- and area- level data is usually subject to different kinds of measurement errors. We propose a multi-level model with level-specific regularizations to overcome these issues and use unit- and area-level data jointly for model-based small area estimation. In the process, we evaluate several mixed-norm regularizations to determine an optimal penalization strategy for a set of potential data constellations. All developed methods are tested within a Monte Carlo simulation study. Further, an empirical application is provided on the example of regional health measurement in Germany. We combine health survey data on the unit-level and aggregated micro census records on the area-level to estimate hypertension prevalence.

E0796: **Integer programming and machine learning for computational statistics**
Presenter: **Ulf Friedrich**, Technical University of Munich, Germany
Co-authors: Jan Pablo Burgard

Optimization problems are omnipresent in modern computational statistics. These typically involve computationally challenging instances and often integrality constraints on some or all of the optimization variables, e.g., to describe (binary) decisions within the model. It is therefore necessary to employ discrete optimization techniques. Methods from Integer Programming are especially relevant because powerful general-purpose solvers are available. However, based on the increasing computational power and improved processing of big data, the classical techniques to solve discrete optimization problems are often replaced by Machine Learning algorithms. While Machine Learning can generally not guarantee that a global optimum is computed, it is often several orders of magnitude faster in practical problem solving, in particular on large instances. We combine Machine Learning and Integer Programming in a single algorithm that is fast and exact. A Branch & Bound framework adapted from Integer Programming is used to control the overall progress. By solving relaxed problems and analyzing the optimality gap, the quality of the solutions is monitored and global optimality can be proved. Simultaneously, fast Machine Learning techniques are used to generate good feasible

solutions as early as possible and reduce the size of the search tree. We present a computational study that shows how this integrated approach is advantageous for large regression problems and related questions.

E0817: **Cross-validation subset selection for regression**
Presenter: **Dennis Kreber**, Trier University, Germany

A linear regression model is considered for which we assume that many of the observed regressors are irrelevant for the prediction. In order to avoid overfitting, we want to conduct a variable selection and only include the true predictors for the least square fitting. The best subset selection gained a lot of interest in recent years for addressing this objective. For this method, a mixed-integer optimization problem is solved which finds the optimal subset not larger than a given natural number k with respect to the in-sample error. In practice, a best subset selection is computed for each k, and the ideal k is then chosen via a validation. We argue that the notion of the best subset selection might be misaligned with the statistical intention. Only the sparsity is selected via a validation whereas the best cardinality-constrained subset is selected in accordance to the training error. We address this issue by proposing a discrete optimization formulation which conducts an in-model cross-validation. The proposed program is only allowed to fit coefficients to training data, but it can choose to switch variables on and off in order to minimize the validation error of the cross-validation. Moreover, we conduct a simulation study and provide evidence that the novel mixed-integer formulation provides more accurate predictions than the best subset selection and other prominent sparse regression methods like Lasso.

E0885: **Small area estimation of transition probabilities for spatial dynamic microsimulation models in socio-economic research**
Presenter: **Simon Schmaus**, Trier University, Germany
Co-authors: Jan Pablo Burgard, Joscha Krause

Dynamic microsimulations allow for the analysis of complex socio-economic systems. First, a synthetic replica of the corresponding system on a micro level as the base population is generated. Next, the replica is stochastically projected into future periods under different scenarios. Comparing the simulation outcomes then provides insights on essential system properties. The projection requires the definition of transition probabilities for every system-intrinsic entity. In order to obtain authentic simulation results, they must reflect the characteristic dynamics of the system. As corresponding probabilities are unknown in practice, they must be estimated from survey data. However, transition probability estimation can be challenging. If the system is spatially segmented into heterogeneous areas, dynamics may vary locally and regional estimation is required. Regional probability estimates may be subject to high uncertainty if the survey data lacks in local observations. We discuss methods of small area estimation and regional benchmarking for transition probability estimation in spatial dynamic microsimulations. The methodology is demonstrated in a simulation study based on a dynamic microsimulation model for care research in Germany.

EO170 Room Court CLUSTERING AND CLASSIFICATION Chair: Maria Brigida Ferraro

E0747: **Clustering objects in three-way proximity data**
Presenter: **Donatella Vicari**, Sapienza University, Italy
Co-authors: Laura Bocci

Three-way two-mode proximity data consist of several symmetric matrices of pairwise proximities between N objects coming from H different subjects (or other data sources such as occasions, experimental conditions, scenarios, time points). In such a context, it is interesting to investigate how the perception or evaluation of the similarities between objects may differ across several subjects. Since the heterogeneity among subjects could entail different classifications of the objects, clustering objects in three-way proximity data is a complex task. Different solutions have been proposed in the statistical literature to deal with the subject heterogeneity. Starting from the INDCLUS (INdividual Differences CLUStering) model and considering that a unique common object clustering is often not enough to account for the subject heterogeneity, a new clustering model is presented where some of the objects (say $N_p \leq N$) belong to common non-overlapping clusters, while the remaining $N - N_p$ objects belong to H different individual (subject-specific) partitions where clusters are linked one-to-one to the common ones. The idea is that some common object clusters exist and form the roots at which the individual clusters intersect accounting for the heterogeneity of the subjects. The model is fitted in a least-squares framework and an efficient Alternating Least Squares algorithm is provided.

E1166: **Fuzzy clustering-based non-linear dimensionality reduction**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

Non-linear dimensionality reduction is a key component of recent machine learning based techniques for large and complex data. This is because such data exist in a mathematically non-linear observational space, and it is difficult to obtain a meaningful and clear latent structure of the observational data with conventional linear dimensional reduction methods. Foremost reasons for this difficulty are the sensitivity of the data in the non-linear observational space and that the simple linear dimensionality reduction techniques lose important sensitive information which is part of the original observation. We propose a new fuzzy clustering-based non-linear dimensionality reduction method to overcome these problems. This method utilizes fuzzy clustering to consider the nonlinearity of the observational space and involves captured features of the nonlinearity to the conventional linear dimensionality reduction method. In addition, by including the difference of fuzzy clustering results over the times from the originally observed data over the times, visualization for the difference of data over times by using the proposed method is presented. Several numerical examples show the better performance of the proposed method.

E1382: **Variable role screening for high-dimensional model-based discriminant analysis**
Presenter: **Michael Fop**, University College Dublin, Ireland
Co-authors: Pierre-Alexandre Mattei, Thomas Brendan Murphy, Charles Bouveyron

Discriminant analysis is a popular supervised classification method used in a variety of fields. Application of this method to high-dimensional data faces major challenges and scalable variable selection plays an important role in the case of sample size much smaller than the number of predictor variables. Sure independence screening is a simple yet effective method to this purpose, consisting of large-scale screening followed by moderate-scale variable selection, with the aim to reduce the high-dimensional problem to a manageable scale. Although allowing for scalable inference, this approach is based on an independence assumption of the predictors which is often too restrictive in practical data analysis and masks the presence of potential redundant variables highly correlated with the actual relevant ones. We discuss a variable screening method which relaxes this independence assumption and where different models are specified according to the roles that each pair of predictor variables takes with respect to the target class variable, either relevant, redundant or uninformative. The approach is specified within a Bayesian framework, where the use of conjugate priors allows for fast and efficient computations of Bayes factors employed to compare the different models. The method is shown in application to data examples and its practical implementation in high-dimensional settings is discussed.

E1640: **An extended k-means clustering procedure with unique factors**
Presenter: **Masamichi Ito**, Osaka University, Japan
Co-authors: Kohei Adachi

A k-means clustering (KMC) procedure is performed for a data matrix of observations by variables with the purpose of classifying the observations into a few clusters. In KMC, the data matrix is modeled as the product of membership and cluster center matrices plus an error matrix. The KMC model is extended by incorporating a unique factor part: we propose a clustering procedure, in which the data matrix is modeled as the sum of an error matrix, the product in the original KMC model, and the unique factor part. This part is the product of a unique factor score matrix and a diagonal matrix, with the former matrix constrained to be column-orthonormal. This orthonormality and the latter matrix being diagonal imply that the unique factors and the variables have a one-to-one correspondence, and each of the factors explains specifically the variation in the

correspondence variables which remains unaccounted for by clusters. Thus, the proposed procedure is useful for analyzing the data set including the variables whose variations are not explained well by clusters. We present an alternating least squares algorithm for the proposed procedure and assess its behaviors numerically.

E1803: **Enhanced algorithms for hierarchical clustering**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Ana Maria Puiu

Two hierarchical clustering approaches are investigated. The first is based on a previously proposed iterative clustering algorithm. It follows a deterministic approach that tries to improve a current clustering, represented as a binary tree. At each step two local moves are performed: reordering and graft, and the best alternative is chosen. A randomized version is proposed such that a non-improving move is accepted with a specified probability. The second approach considers a combinatorial and two randomized clustering procedures that are based on the Ward method. The Combinatorial Hierarchical Clustering differs from the original Ward method by the fact that at each merging step the execution is divided into two new execution threads, one in which the best cluster pair is merged, and a second one in which the second best cluster pair is merged. The execution plan can be visualized as a binary tree in which each node has two edges representing the current choice (best or second best) and the child nodes represent the partial solutions generated so far. A Restricted Randomized Hierarchical Clustering differs from the original Ward method by the fact that at each merging step the second best clustering pair is merged instead of the optimal one with a specified probability. A generalization of this method that randomly chooses from all merging possibilities is also introduced. Experimental results are performed and analyzed.

EO260 Room Jessel RECENT ADVANCES IN TIME SERIES ANALYSIS Chair: Masayuki Hirukawa

E0557: **Estimating average derivative with multiple integrated regressors**
Presenter: **Anurag Banerjee**, Durham University, United Kingdom

The model $y_t = m(\mathbf{x}_t) + u_t$ is considered, where the covariates $\mathbf{x}_t$ are $d-$dimensional integrated variables ($d \geq 3$). The equally weighted average derivative (AD) of the regression function m within a bounded box $\mathcal{K}$ is defined as $(1/|\mathcal{K}|)\int_{\mathcal{K}} D[m(z)]dz$. The AD is then estimated using piecewise local linear regression. We study the asymptotic distribution of this estimator. The results indicate that the ADE converges at the rate $T^{-1/3}$ when $m(.)$ is non-linear and T^{-1} if $m(.)$ is linear. We provide a randomised algorithm to estimate the AD. Using Monte-Carlo simulation experiments, we investigate the small sample properties of our estimator.

E1447: **Testing linear cointegration against smooth transition cointegration**
Presenter: **Martin Wagner**, University of Klagenfurt, Austria
Co-authors: Oliver Stypka

Simple tests are developed for the null hypothesis of linear cointegration against the alternative of smooth transition cointegration. The test statistics are based on the fully modified or integrated modified OLS estimators suitably modified to Taylor approximations of smooth transition functions. This necessitates the adaptation of the above estimation approaches to models including cross-products of integrated regressors. As transition variable we consider integrated variables and time. For the integrated modified OLS based test we additionally develop fixed-b inference. The properties of the tests are evaluated with a simulation study and compared to a previous test. Finally, we apply our tests to investigate money demand for eight countries or areas, inter alia the Euro Money Area with data from 1995Q1 and the USA with data from 1964Q1. For interest rate and time as transition variable there is strong indication against the null of linearity.

E1501: **Inference on time series with systematically missing data**
Presenter: **Fulvia Marotta**, Queen Mary University of London, United Kingdom

Unevenly spaced observations is a fundamental issue in econometric time series analysis. There has been considerable focus on parametric estimation methods permitting systematically missing data or observations missing at random times. We suggest a new model fitting approach for such data. The basic idea is to correct the periodogram for missing data so that it enables standard parametric Whittle estimation. After missing data have been accounted for, the dynamics parameters of a time series can be estimated with parametric rate and confidence intervals constructed. The main advantage of such modelling is its simplicity and easiness of use in applications. We investigate performance of the method with simulated and empirical data.

E1521: **Investigation of statistical property in residual series after detrending**
Presenter: **Hiroko Solvang**, Institute of Marine Research, Norway

During the last decade, global warming has led to changes in spatial distribution of plankton and fish in marine ecosystem. There is an urgent need for a holistic Ecosystem Based management of the ecosystem, taking the rapid climate driven changes in the natural systems. A simulation-based study for future scenario is necessary to evaluate climate change impact on biological ecosystem component, which especially affects fish stock dynamics under different management strategies. To generate realistic simulation data, the information extracted from time series observation recorded by annual survey should be involved. The data present long-run movements caused by temperature and short/middle cyclic terms in temporal changes in ecosystem. The stationary statistical property of the data is investigated by applying another statistical analysis to short/middle cyclic terms. We propose a systematic statistical procedure, integrating time series decomposition and robust statistical testing procedures. This procedure estimates the trend component corresponding to long-run movements and identifies the statistical property for the detrended component. Based on the simulation data, future scenarios of complex temperature variation and cyclic fluctuation may be predicted.

E0266: **Identifying market manipulation & abusive trading using anomaly detection techniques**
Presenter: **Robert James**, The University of Sydney, Australia
Co-authors: Artem Prokhorov, Henry Leung

A novel semi-supervised procedure is presented to detect instances of intraday abusive trading in financial markets, addressing the limitations present in existing rule-based expert surveillance systems, which are pervasive within the industry. It is considered that abusive trading produces highly abnormal patterns in the time series of limit order book activity, making such abusive activity detectable even in the absence of explicit assumptions regarding its form. We employ a state-of-the-art optimized implementation of the K-Nearest Neighbour Dynamic Time Warping algorithm to compute the similarity between multivariate time series sub-sequences of trading activity. A threshold defining the boundary between normal and abusive activity is constructed by applying univariate extreme value theory to the set of DTW similarity scores observed under estimated normal trading conditions. Using real world, tick-by-tick transaction data provided by a global investment bank we highlight the utility of the procedure in identifying instances of insider trading and demonstrate its competitiveness with respect to several benchmark algorithms used in related literature.

EO332 Room MAL 152 BANDWIDTH SELECTION FOR KERNEL ESTIMATION Chair: Maria-Dolores Martinez-Miranda

E0285: **Subsampling-extrapolation bandwidth selection in bivariate kernel density estimation**
Presenter: **Qing Wang**, Wellesley College, United States
Co-authors: Adriano Zambom

The focus is on bivariate kernel density estimation that bridges the gap between univariate and multivariate applications. We propose a subsampling-

extrapolation bandwidth matrix selector that improves the reliability of the conventional cross-validation method. The proposed procedure combines a U-statistic expression of the mean integrated squared error and asymptotic theory, and can be used in both cases of diagonal and unconstrained bandwidth matrix. In the subsampling stage, one takes advantage of the reduced variability of estimating the bandwidth matrix at a smaller subsample size m ($m < n$); in the extrapolation stage, a simple linear extrapolation is used to remove the incurred bias. Simulation studies reveal that the proposed method reduces the variability of the cross-validation method by about 50% and achieves an expected integrated squared error that is up to 30% smaller than that of the benchmark cross-validation. It shows comparable or improved performance compared to other competitors across six distributions in terms of the expected integrated squared error. We prove that the components of the selected bivariate bandwidth matrix have an asymptotic multivariate normal distribution, and also present the relative rate of convergence of the proposed bandwidth selector.

E1052: **A new framework for kernel intensity estimation in point processes using covariates**
Presenter: **Maria Isabel Borrajo**, Universidade de Santiago de Compostela, Spain
Co-authors: Wenceslao Gonzalez-Manteiga, Maria-Dolores Martinez-Miranda

The bias-variance trade-off for inhomogeneous point processes with covariates is addresses theoretically and empirically. A consistent kernel estimator for the first-order intensity function is constructed, using a convenient relationship between the intensity and the density of events location. The asymptotic bias and variance of the estimator are derived and hence the expression of its infeasible optimal bandwidth. Three data-driven bandwidth selectors are proposed to estimate the optimal bandwidth. One of them is based on a new smooth bootstrap proposal that is proved to be consistent under a Poisson assumption. The other two are a rule-of-thumb method based on assuming normality, and a simple non-model-based approach. An extensive simulation study is accomplished considering Poisson and non-Poisson scenarios, and including a comparison with other competitors in the literature. The practicality of the new proposals is shown through an application to real data about wildfires in Canada, using meteorological variables as covariates.

E0316: **Data-driven bandwidth selection for recursive kernel density estimators under double truncation**
Presenter: **Yousri Slaoui**, University of Poitiers, France

A data-driven bandwidth selection procedure is proposed for the recursive kernel density estimation under double truncation. We show that using the selected bandwidth and a special stepsize, our proposed recursive estimators outperform the nonrecursive one in terms of estimation error in many situations. We corroborate these theoretical results through a simulation study. The proposed estimators are then applied to data on the luminosity of quasars in astronomy.

E0808: **Bandwidth selection under random fields with short memory**
Presenter: **Bastian Schaefer**, University of Paderborn, Germany

The problem of bandwidth selection in semiparametric spatial models under dependent errors is studied. We use a spatial representation of high-frequency financial data on a lattice and aim at estimation of a non-stationary regression surface and a stationary component with short memory in all dimensions, e.g. volatility or other risk measures. The non-stationary mean surface function is estimated by a nonparametric method, the demeaned residuals are modeled as a random field which is estimated by parametric methods. We propose bandwidth selectors for estimation of the regression surface under different random field representations of the error terms. Optimal bandwidths are chosen by an iterative plug-in algorithm to minimize the mean integrated squared error of the nonparametric estimator. An implementation into an R package is given and the properties of the estimator are assessed under a simulation study.

E0956: **Creating a safe space for bias correction in kernel hazard estimation**
Presenter: **Maria-Dolores Martinez-Miranda**, Universidad de Granada, Spain
Co-authors: Vali Asimit, Maria Luz Gamiz, Jens Perch Nielsen

Bias-correction methods in kernel estimation are often knocked down by an explosion in variance. Considering the simplest situation of kernel hazard estimation we propose a new algorithm to reduce the bias with no cost in terms of variance. We start the algorithm in a not so complex model (rather large bias but small variance) and move with a safety net towards a more complex model (reducing the bias). We move in a safe way smoothing the slope of the ride. The end result is that we get the variance from our starting place and the bias from our ending place. The algorithm is therefore a safe ride from a non-complex world towards complexity where we get the best of both worlds. Simulation experiments and asymptotic theory support our proposal.

EO350 Room Senate STATISTICAL ADVANCES IN EXTREMES AND RISK MANAGEMENT Chair: Gilles Stupfler

E0594: **Multivariate geometric expectiles and range value-at-risk**
Presenter: **Melina Mailhot**, Concordia, Canada
Co-authors: Marius Hofert, Klaus Herrmann

Geometric generalizations of expectiles and Range Value-at-Risk for d-dimensional multivariate distribution functions will be introduced. Multivariate geometric expectiles are unique solutions to a convex risk minimization problem and are given by d-dimensional vectors. Multivariate geometric Range Value-at-Risk is a risk measure considering tail events, which has TVaR as a special case. They are well behaved under common data transformations. Properties and highlights on the influence of varying margins and dependence structures will be presented.

E0685: **Nonparametric extreme conditional expectile estimation**
Presenter: **Antoine Usseglio-Carleve**, Inria, France
Co-authors: Stephane Girard, Gilles Stupfler

Expectiles and quantiles can both be defined as the solution of minimization problems. Contrary to quantiles though, expectiles are determined by tail expectations rather than tail probabilities, and define a coherent risk measure. For these two reasons in particular, expectiles have recently started to be considered as serious candidates to become standard tools in actuarial and financial risk management. However, expectiles and their sample versions do not benefit from a simple explicit form, making their analysis significantly harder than that of quantiles and order statistics. This difficulty is compounded when one wishes to integrate auxiliary information about the phenomenon of interest through a finite-dimensional covariate, in which case the problem becomes the estimation of conditional expectiles. We propose nonparametric estimators of extreme conditional expectiles based on kernel smoothing techniques. We analyze the asymptotic properties of our estimators in the context of conditional heavy-tailed distributions. Applications to simulated and real data are provided.

E0992: **A novel GARCH-EVT approach dealing with bias and heteroscedasticity for extreme risk estimations**
Presenter: **Hibiki Kaibuchi**, SOKENDAI The Graduate University of Advanced Studies, Japan
Co-authors: Gilles Stupfler, Yoshinori Kawasaki

Extreme Value Theory (EVT) has not yet emerged as a dominating tool in financial risk management, i.e. extreme risk estimations. This is due to the time-varying volatility of financial time series. In order to overcome this problem, the two-step GARCH-EVT approach was introduced. It should be noted that one drawback of this methodology is that the correction of bias is not thoroughly considered. We propose a new way, as far as we aware, to estimate conditional VaR considering both bias correction and volatility background based on original GARCH-EVT approach. For that, we: (i) pre-whiten the financial time series with a GARCH(1,1) model for forecasting volatility; (ii) apply the semi-parametric bias-corrected tail estimators to standardized residuals from the GARCH analysis. We also consider the extension of extreme expectile estimation to a time dynamic setting. The results are illustrated on simulated data and on a financial real dataset.

E1205: **On the strong convergence of sample maxima, with applications to asymptotic statistical theory for extremes**
Presenter: **Stefano Rizzelli**, EPFL, Switzerland
Co-authors: Michael Falk, Simone Padoan

It is well known and readily seen that the maximum of n independent and uniformly on $[0,1]$ distributed random variables, suitably standardized, converges in total variation distance, as n increases, to the standard negative exponential distribution. We extend this result to higher dimensions by considering copulas. We show that the strong convergence result holds for copulas that are in a differential neighbourhood of a multivariate generalized Pareto copula. We provide sufficient conditions for such a property to be satisfied, under a suitable assumption on the underlying extreme-value copula structure. Sklar's theorem then implies convergence in variational distance of the maximum of n independent and identically distributed random vectors with (arbitrary) common distribution function and their appropriately normalized version. We illustrate how these convergence results can be exploited to establish the almost-sure consistency of inferential procedures for max-stable models, using sample maxima.

E1256: **Extreme expectile estimation for heavy-tailed time series**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Gilles Stupfler

Extreme quantiles estimation have been widely discussed in the literature, since that for instance the value at risk is an important "measure" for the risk quantification in many applied fields such as in finance and insurance. In finance, real-life time series often reveal a certain degree of dependence through time, therefore, in the last years, a big effort has been devoted in deriving inferential methods for extreme quantiles that accommodate for the temporal dependence. The tail expectile is a coherent risk measure, therefore it provides an appealing alternative to the value at risk. Recently, several estimation results for extreme expectiles are emerging, under the assumption that the data are independent. We investigate the behaviour (asymptotic properties and finite sample performances) of some expectiles estimators based on large observations of stationary time series, under mild assumptions on the serial dependence.

EO793 Room CLO 203 HUMAN MICROBIOME RESEARCH: NEW DESIGNS AND STATISTICAL METHODS Chair: Ekaterina Smirnova

E1939: **Microbiome data: Current challenges and opportunities**
Presenter: **Ekaterina Smirnova**, Virginia Commonwealth University, United States

The composition of microbial species in a human body is essential for maintaining human health, and it is associated with a number of diseases including obesity, bowel inflammatory disease, and bacterial vaginosis. Over the last decade, microbiome data analysis almost entirely shifted towards using samples taken directly from various sites of human body and to explore a large number of microbes using 16S or whole metagenome sequencing. With the growth and success of the microbiomics field, the size and complexity, and availability of the microbiome data in any given experiment have increased exponentially. We discuss several challenges that currently remain in microbiome research and directions towards resolving them. These include: 1) the lack of reproducibility in experiments; 2) statistical challenges in designing and analyzing longitudinal studies; and 3) opportunities for working with publicly available data sets. In particular, we discuss the variation in microbial samples between sampling processing methods and testing location discovered by the Microbiome Quality Control Project (MBQC), as well as approaches towards alleviating this technical variation.

E1517: **Compositional mediation model for binary outcome: Application to microbiome samples**
Presenter: **Michael Sohn**, University of Rochester, United States

The delicate balance of the microbiome is implicated in our health and is shaped by external factors. Therefore, understanding the mediating role of the microbiome in linking external factors and our health conditions is crucial to translate the microbiome research into therapeutic and preventative applications. We introduce a sparse compositional mediation model for binary outcomes under potential outcomes framework to estimate and test the causal mediation effect utilizing the compositional algebra defined in the simplex space and a linear zero-sum constraint on regression parameters.

E1817: **Accounting for asymmetry and batch effects in meta-transcriptomics**
Presenter: **Gregory Gloor**, University of Western Ontario, Canada

The ALDEx2 tool generates a probabilistic model of high throughput sequencing count data and manipulates that model using the tools and rules of compositional data analysis. Many biological experiments are conducted using data gathered via high throughput sequencing. These data are widely regarded as delivering counts per feature in each sample. However, the counts are more properly regarded as relative abundances (compositions) because the instrument imposes a restriction on the upper limit on total counts. Thus, increases in counts of one sample (or feature) must be compensated by a decrease in the counts of another sample or feature. ALDEx2 has recently incorporated general linear models and the ability to use asymmetric datasets in the analysis of high throughput sequencing datasets. We use three different vaginal meta-RNA-seq datasets from different labs and show that the combination of the GLM and asymmetry correction allow a principled meta-analysis of the joint datesets. We find that the most effective approach is to hold 'housekeeping' functions as constant as possible and to determine the change in expression relative to those functions. Confounders such as read length, sequencing platform and read depth are not observed to have a significant effect.

E1921: **Adapting methods of gene set enrichment analysis to study of the human microbiome**
Presenter: **Levi Waldron**, CUNY Graduate School of Public Health and Health Policy, United States

Despite advances in methods for differential abundance analysis in microbiome studies, biological interpretation of such results remains challenging. In a recent study of the oral microbiome in oral rinse specimens of heavy smokers, never smokers, and individuals likely exposed to second-hand smoke from the New York City Health and Nutrition Examination Study, we observed strong evidence of community-wide microbiome shifts correlated with primary and second-hand tobacco exposure. However, mechanistic interpretation of these shifts was challenging. To test the hypothesis that smoking affects oral microbiota based on their oxygen requirements, we annotated all bacterial genera observed in our study as aerobic, anaerobic, or facultatively anaerobic, creating corresponding microb sets for microbe set enrichment analysis. This analysis identified depletion of aerobic genera in smokers to be highly significant ($p < 0.001$, permutation test) which, combined with more traditional methods of causal inference, strongly suggest differentially harmful effects of cigarette smoke on the aerobic oral microbiome, rather than effects of unmeasured confounding. Our current efforts are discussed aiming at curating a broad base of microbial signatures to enable the development of methods for systematic microbe set enrichment analysis. New challenges and areas of research for statistical methods development are presented.

E1945: **Experimental and computational frameworks for microbiome data analysis assessment**
Presenter: **Hector Corrada Bravo**, University of Maryland, College Park, United States

Analysis of 16S rRNA marker-gene surveys may be performed by a variety of bioinformatic pipelines and downstream analysis methods. However, appropriate assessment datasets and statistics are needed as there is limited guidance to decide between available analysis methods. Mixtures of environmental samples are useful for assessment as they provide values calculated from measurements of the unmixed samples and the mixture design that can be compared to values recovered by each bioinformatic method. We present an assessment framework for 16S rRNA sequencing analysis methods based on a two-sample titration mixture dataset and metrics to evaluate OTU count table characteristics. The qualitative assessment evaluates feature presence/absence exploiting features only present in unmixed samples or titrations by testing if random sampling can explain their observed relative abundance. The quantitative assessment evaluates how well relative and differential abundance values agree with values expected from the mixture design. To demonstrate the assessment framework, we present results assessing estimates of abundance, differen-

tial abundance and beta diversity from count tables generated using three of the most-commonly used bioinformatic pipelines for this analysis. The developed assessment framework serves as a valuable community resource for assessing 16S rRNA marker-gene survey bioinformatic methods.

EO354 Room CLO 204 RECENT DEVELOPMENT IN SCIENTIFIC AND CLINICAL STUDIES OF THE BRAIN Chair: Guofen Yan

E0200: **A Bayesian approach for mapping epileptic brain networks**
Presenter: **Tingting Zhang**, University of Virginia, United States

The brain regions and the influences exerted by each region over another, called directional connectivity, form a directional network. We study normal and abnormal directional brain networks of epileptic patients using their intracranial EEG (iEEG) data, which are multivariate time series recordings of many small brain regions. We propose a high-dimensional state-space multivariate autoregression (SSMAR) for iEEG data to model the brain as a dynamics system. We assume that the underlying brain network has a cluster structure and develop a Bayesian framework to estimate the proposed high-dimensional model, identify clusters of densely-connected brain regions, and map epileptic patients' brain networks in different seizure stages. We show that the new method is robust to various deviations from the model assumptions, low iEEG sampling frequency, and data noise. Applying the developed Bayesian approach to an epileptic patient's iEEG data, we reveal the patient's network changes at the seizure onset and the unique connectivity of the seizure onset zone (SOZ), where seizures start and spread to other normal regions. Using this network result, our method has the potential to assist clinicians to localize the SOZ.

E0306: **Interpretable principal components analysis for multilevel multivariate functional data**
Presenter: **Robert Krafty**, University of Pittsburgh, United States
Co-authors: Jun Zhang, Greg Siegle

Many studies collect functional data from multiple subjects that have both multilevel and multivariate structures. An example of such data comes from popular neuroscience experiments where participants' brain activity is recorded using modalities such as EEG or fMRI and summarized as power within multiple time-varying frequency bands at multiple brain regions. An important question is summarizing the joint variation across multiple frequency bands for both whole-brain variability between subjects, as well as location-variation within subjects. We discuss a novel approach to conducting interpretable principal components analysis on multilevel multivariate functional data that decomposes total variation into subject-level and replicate-within-subject-level (i.e. electrode-level) variation, and provides interpretable components that can be both sparse among variates (e.g. frequency bands) and have localized support over time within each frequency band. The sparsity and localization of components is achieved by solving an innovative rank-one based convex optimization problem with block Frobenius and matrix L_1-norm based penalties. The method is used to analyze data from a study to better understand blunted affect, revealing new neurophysiological insights into how subject- and electrode-level brain activity are connected to the phenomenon of trauma patients "shutting down" when presented with emotional information.

E0994: **Learning structural connectivity of the brain from imaging data**
Presenter: **Jie Peng**, University of California Davis, United States

In recent years, there has been an explosion of multi-modal brain imaging data due to increasing interests in understanding functional and structural connectivity of the brain. One of the imaging technologies – diffusion MRI (D-MRI) – is an in vivo and non invasive technology that uses water diffusion as a proxy to probe architecture of biological tissues. D-MRI has been widely used in white matter fiber tracts reconstruction as well as many clinical applications including neurodegenerative diseases such as Alzheimer's disease. We will discuss how to extract structural connectivity information from D-MRI data. We will explore various models and challenges through both synthetic experiments and applications on data from large brain imaging consortium such as the Human Connectome Project (HCP) and Alzheimer's Disease Neuroimaging Initiative (ADNI).

E1108: **Identification of spreading depolarization based on potassium and glucose levels**
Presenter: **Satish Iyengar**, University of Pittsburgh, United States

The US is experiencing a large incidence of traumatic brain injury with 1.7 million per year and over 50,000 deaths. Patients with severe TBI require neurosurgery followed by intensive care, with a difficult road to recovery, often with poor outcomes such as severe disability, a vegetative state, or death. Secondary brain injury, or the expansion of the primary lesion into the vulnerable penumbra, is a major contributor to poor outcomes. There are currently no technologies to diagnose secondary injury, so it is untreated. Spreading depolarization triggers a metabolic crisis and is correlated with secondary injury. We will describe a preclinical study of an animal model to expand our knowledge of strategies to mitigate the detrimental consequences of localized tissue damage when probes are inserted into brain tissue. Specifically, we study rats that are subject to a controlled cortical impact, after which glucose and potassium traces are collected. We will describe the use of generalized additive models with autoregression on these bivariate traces to develop methods to identify SDs.

E1901: **Soft tensor regression**
Presenter: **Georgia Papadogeorgou**, Duke University, United States
Co-authors: Zhengwu Zhang, David Dunson

Statistical methods relating tensor predictors to scalar outcomes in a regression model generally vectorize the tensor predictor and estimate the coefficients of its entries employing some form of regularization, use summaries of the tensor covariate, or use a low dimensional approximation of the coefficient tensor. However, low rank approximations of the coefficient tensor can suffer if the true rank of the tensor is not small. We propose a tensor regression framework which assumes a soft version of the parallel factors (PARAFAC) approximation. In contrast to classic PARAFAC, where each entry of the coefficient tensor is the sum of products of row-specific contributions across the tensor modes, the soft tensor regression (Softer) framework allows the row-specific contributions to vary around an overall mean. A Bayesian approach to inference is followed, and it is shown that softening the PARAFAC increases model flexibility, leads to improved estimation of coefficient tensors, and more accurate predictions, even for a low approximation rank. In the context of the motivating application, Softer is adapted to symmetric and semi-symmetric tensor predictors and is used to analyze the relationship between brain network characteristics and human traits.

EO835 Room MAL 252 RESTRICTED PARAMETERS INFERENCE AND SHRINKAGE ESTIMATORS Chair: Genso-Y.-T. Watanabe-Chang

E0429: **Shrinkage priors on complex statistical manifolds**
Presenter: **Hidemasa Oda**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

It is difficult to define the best properties that statistical manifolds should possess. One of the good properties we believe statistical manifolds have is their Kaehler structure. Complex-valued stochastic processes are useful models for parametrizing complex or bivariate signals. We expanded the theory of α-geometry of real time series for complex time series. For most of the parts, generalized results of real time series for complex time series are obtained. We are interested in particular in the case when the information manifold is a Kaehler manifold. It has been previously shown that the information geometry of complex time series is Kaehler. We further investigate the structure of the complex autoregressive models and its positive superharmonic priors. We expect that Ricci-free α-Kaehler structure has an important role in the theory of complex information geometry. We will discuss the application of α-Kaehler geometry for general complex linear systems.

E0436: **Pitman closeness domination in predictive density estimation for two ordered normal means under alpha-divergence loss**
Presenter: **Genso-Y.-T. Watanabe-Chang**, Mejiro University, Japan
Co-authors: Nobuo Shinozaki, William Strawderman

Pitman closeness domination in predictive density estimation problems is considered when the underlying loss metric is the alpha-divergence, D(alpha). The considered underlying distributions are normal location-scale models, including the distribution of the observables, the distribution of the variable whose density is to be predicted, and the estimated predictive density which will be taken to be of the plug-in type. The scales may be known or unknown. A general expression for the alpha-divergence loss in this set-up has been previously derived. It has been shown that it is a concave monotone function of the quadratic loss, and also a function of the variances (predicand, and plug-in). We demonstrate the D(alpha)-Pitman closeness domination of certain plug-in predictive densities over others for the entire class of metrics simultaneously when modified Pitman closeness domination holds in the related problem of estimating the mean. We also establish D(alpha)-Pitman closeness results for certain generalized Bayesian (best invariant) predictive density estimators. Examples of D(alpha)-Pitman closeness domination presented relate to the problem of estimating the predictive density of the variable with the larger mean. We also consider the case of two ordered normal means with a known covariance matrix.

E0458: **On estimation and prediction for high-dimensional Poisson models with quasi zero inflation**
Presenter: **Keisuke Yano**, The University of Tokyo, Japan

The problem of estimating and predicting high-dimensional Poisson models with quasi zero inflation is discussed. In analyzing count data, exact or quasi zero inflation often appears. Inference for zero-inflated count data has gathered much attention in a wide range of applied areas. We present results for both estimation and prediction of Poisson models with zero inflation. We investigate the asymptotic minimax risk that indicates a clear correspondence between Gaussian and Poisson models with sparsity. An easy-implemented asymptotically minimax estimator is constructed using a spike-and-slab prior with a polynomially-decaying slab prior. We also investigate asymptotically minimax predictive densities that are adaptive to an unknown sparsity level. To achieve this result, we have employed a new idea of calibrating the scale of the slab prior according to the minimax risk.

E0479: **On prediction and shrinkage estimation for balanced loss functions**
Presenter: **Eric Marchand**, Universite de Sherbrooke, Canada
Co-authors: William Strawderman

The estimation of a multivariate mean θ is considered under natural modifications of balanced loss function of the form: (i) $\omega\rho(\|\delta-\delta_0\|^2)+(1-\omega)\rho(\|\delta-\theta\|^2)$, and (ii) $\ell\left(\omega\|\delta-\delta_0\|^2+(1-\omega)\|\delta-\theta\|^2\right)$, where δ_0 is a target estimator of $\gamma(\theta)$. After briefly reviewing known results for original balanced loss with identity ρ or ℓ, we provide, for increasing and concave ρ and ℓ which also satisfy a completely monotone property, Baranchik-type estimators of θ which dominate the benchmark $\delta_0(X)=X$ for X either distributed as multivariate normal or as a scale mixture of normals. Implications are given with respect to model robustness and simultaneous dominance with respect to either ρ or ℓ. Finally, we present a framework for predictive density estimation under balanced loss functions, we describe Bayesian representations and discuss frequentist performance of various predictive density estimators.

E0583: **Shrinkage estimation of the mean of high-dimensional normal distribution**
Presenter: **Ryota Yuasa**, The University of Tokyo, Japan
Co-authors: Tatsuya Kubokawa

The problem of estimating the mean matrix of the multivariate normal distribution in high dimensional setting is addressed. Efron-Morris-type estimators with ridge-type inverse matrices are considered. The proposal to estimate optimal weights is based on the minimization of the risk function under a quadratic loss. The proposed estimators are derived by using Stein's identity. It is shown that the proposed estimators have minimaxity. Furthermore, by using Random Matrix Theory, it can be shown that proposed estimators are optimal from the viewpoint of the asymptotic minimization of the loss function when a prior distribution is assumed for the true mean. Numerical experiments are conducted to confirm the performance of the estimators, which are compared with Efron-Morris, James-Stein and singular value shrinkage type estimators. The proposed estimators provide better accuracy than the others, especially when both n and p are large.

EO428 Room MAL 253 OCEANS OF DATA: HIGH-DIMENSIONAL STATISTICS FOR MARINE DATA ANALYSIS Chair: Christian L. Mueller

E1249: **Marine data mining with CMAP using R**
Presenter: **Aditya Mishra**, Flatiron Institute, Simons Foundation, United States
Co-authors: Christian L. Mueller, Jacob Bien, Sangwon Hyun

Recent advances in experimental techniques and scientific instruments have enabled the collection of biological, biogeochemical, and imaging data of the ocean on a global scale. The Simons CMAP, a currently developed large-scale open-access marine database, hosts a multitude of such marine datasets, including remote-sensing satellite observations, large-scale integrated in-situ biogeochemical cruise measurements, amplicon sequencing data, and complex synthetic ocean simulation data. To facilitate easy access to these rich data sets for statisticians and data scientists, we have developed cmap4r, an R package that enables downloading, analyzing, and visualizing datasets from the Simons CMAP in a fast and structured manner. Integrated analysis of marine data is challenging due to several factors, including the presence of outliers, missing entries, different spatial and temporal resolutions, spatiotemporal dependencies, high dimensionality, and for amplicon sequencing data, the absence of absolute species abundance measurements due to experimental limitations. This presents a unique opportunity for both the development and the application of novel statistical methods for marine data analysis. Using cmap4r as primary access point to the database, we highlight two novel statistical analysis examples where we have developed high-dimensional statistical techniques to relate microbial species abundances, marine environmental factors, and primary productivity in the ocean.

E0997: **Structured feature aggregation in regression with microbiome data**
Presenter: **Jacob Bien**, University of Southern California, United States
Co-authors: Christian L. Mueller, Xiaohan Yan

Microbiome data allow marine ecologists to understand the composition of microbes across time and space in the ocean. Much statistical work has focused on two technical challenges for the analysis of microbiome data: (i) it is high-dimensional, i.e., there are a large number of microbes and (ii) it is not meaningful to directly compare the raw absolute abundances measured, which means compositional data methods are used. The focus is on yet another major challenge, which has received far less attention than the other two: microbiome data has a high degree of sparsity, i.e., the vast majority of microbes measured are generally present in only a few samples. Researchers generally take ad hoc approaches such as filtering out rare microbes or manually aggregating them to the genus or family level. We propose instead a principled regression framework, which we demonstrate addresses all three of these challenges while also providing easily interpretable models.

E1317: **Joint modeling of continuous flow cytometry data with environmental covariates**
Presenter: **Sangwon Hyun**, University of Southern California, United States
Co-authors: Jacob Bien, Francois Ribalet, Mattias Cape

Flow cytometry data collected in the ocean can give valuable insight into the composition and dynamics of phytoplankton populations. We present a novel method for modeling time-varying flow cytometry data conditional on a large number of environmental covariates. We develop a novel mixture of multivariate sparse regressions model that can simultaneously estimate and identify the important covariates for each phytoplankton population. The method ties covariates to both the flow cytometry population centers as well as the relative abundances of these populations. The approach involves a lasso-penalized expectation-maximization procedure with additional convex constraints to facilitate interpretation of the

estimated model. We apply the method to continuous-time flow cytometry data measured from the ocean, on a ship near Honolulu traveling from warmer, nutrient-sparse subtropical waters to cooler, more productive waters. The method provides a powerful framework for developing a fine-grained understanding of the environmental drivers of phytoplankton populations in the ocean.

E0928: **A trait-based taxonomy for phytoplankton biomass modeling and prediction**
Presenter: **Crispin Mutshinda**, Dalhousie University, Canada
Co-authors: Andrew Irwin, Zoe Finkel

A Bayesian model is developed for phytoplankton biomass dynamics and estimate trait values of phytoplankton taxa in the Western English Channel using 7 years of weekly data recorded at Station L4. In addition to the usual classification of species into diatom and dinoflagellate functional types, we create and evaluate an alternative trait-based grouping. While our ultimate interest is in biomass modeling and projection, species-level biomass data are fraught with missing values due either to actual absences or abundances below detection thresholds. Consequently, we initially rely on occurrence data, which are exempt from these issues. We analyze the factors that determine the presence-absence of individual phytoplankton species and subsequently apply a clustering algorithm to these traits to define a new taxonomy. We contrast the performance of the biomass dynamics model under three different groupings namely, (i) our trait-based taxonomy, (ii) the null clustering of identical structure where cluster memberships are randomly assigned to species, and (ii) the functional type taxonomy, with regard to the quality of cluster-specific trait value estimates and the model skill for total biomass prediction.

E1167: **Inference of ecological networks from a sparse ocean dataset**
Presenter: **Joseph Siddons**, Dalhousie University, Canada
Co-authors: Andrew Irwin, Zoe Finkel

The continuous plankton recorder (CPR) survey is a large, multi-decadal, plankton monitoring programme spanning the North Atlantic, and is a powerful tool for investigating the effects of climate change. Using a network inference approach, including sparse neighbourhood selection and correlation analysis, we build ecological interaction networks to investigate the planktonic community structure. We are particularly interested in observing trophic level interactions (such as grazing), and the evolution of communities over time (e.g. between seasons or decades). We attempt to account for the environmental predictors in order to separate the relationships that are driven by niche effects from those that arise as a consequence of biological forcing. A difficulty of the CPR dataset is that it is heavily zero-inflated. Consequently, we need to make use of appropriate correlation estimators for sparse data. We have identified a number of species that are important for community structure, acting as either hub species, or as a bridge between clusters of taxonomic groups. Results indicate that species within plankton taxonomic groups are more positively correlated than species between groups (in particular zooplankton-phytoplankton relationships).

EO464 Room SH349 STATISTICS IN SPORT Chair: Marialuisa Restaino

E0489: **Cultural diversity and team performance in the Italian Serie A**
Presenter: **Francesco Addesa**, Leeds Beckett University, United Kingdom

Cultural diversity features prominently in management studies. A diverse range of skills and perspectives can produce innovation and a greater variety of solutions to day to day problems. At the same time, however, the same heterogeneous approaches and experiences can result in communication and coordination problems, lack of trust and intra/intergroup conflict. Two measures of diversity, indices of fractionalization and polarization, are computed on the basis of a newly constructed dataset on team composition and performance for 30 teams, 1,409 players and 2,279 matches in the Italian Serie A with information on the players place of origin, talent, position, demographics, manager experience and other factors. Only the fractionalization index is found to have a strong and persistent negative effect on game scores and player objective performance ratings. These results provide new important insights into the drawbacks of cultural heterogeneity in the workplace and calls for more in-depth analyses of the nexus diversity-performances on team performance.

E1139: **Injury forecasting in sports using artificial intelligence**
Presenter: **Luca Pappalardo**, ISTI-CNR, Italy
Co-authors: Alessio Rossi, Paolo Cintia

Injuries have a great impact on professional athletes and clubs, due to their large influence on team performance and the considerable costs of rehabilitation for players. Existing studies in the literature provide just a preliminary understanding of which factors mostly affect injury risk, while an evaluation of the potential of statistical models in forecasting injuries is still missing. We propose an approach to injury forecasting in professional sports that is based on artificial intelligence. By using GPS tracking technology, we collect data describing the training workload of players in a professional soccer club during two seasons. We then construct an injury forecaster and show that its accuracy outperforms existing methods for injury risk assessment currently used by professional clubs. Our approach opens a novel perspective on injury prevention, providing a set of simple and practical rules for evaluating and interpreting the complex relations between injury risk and training performance in professional sports.

E1199: **Two decompositions of the Pietra index with applications to Italian professional football teams**
Presenter: **Francesco Porro**, Universita degli Studi di Milano-Bicocca, Italy
Co-authors: Mariangela Zenga

Two innovative procedures for the decomposition of the Pietra inequality index are proposed. Both are based on a two-step approach, already successfully applied in the literature to decompose other indexes. The first procedure allows the decomposition of the index by sources, while the second one provides the decomposition by subpopulations. A very important advantage of these two new procedures is that the first one allows us to assess the relevance of each source, while the second one provides the contribution due to each subpopulation. The "classical" decomposition of the Pietra index in the Within and the Between components can be easily obtained as a special case of the proposed decomposition procedure by subpopulations. Beyond the methodological details, two applications with real data regarding the professional football teams in Italy are illustrated.

E1257: **Weighted ELO rating predictions in tennis**
Presenter: **Vincenzo Candila**, University of Salerno, Italy
Co-authors: Luca De Angelis, Giovanni Angelini

Several methods are available in literature for estimating the probability of winning in tennis, such as the regression-based, point-based and paired-comparison approaches, for instance. Among these latter, the ELO rating method plays a prominent role. Originally applied to tennis by the data journalists of FiveThirtyEight.com, the ELO rating method estimates the strength of each player on the basis of the last match in order to predict the probability of winning for the upcoming match. Notwithstanding its widely recognized merits in terms of ease of reproducibility and good performances, the ELO rating system does not take into account the number of games won by each player in the last match(es). The aim is to investigate the profitability of a variant of the standard ELO rating method, where also the games of the last match(es) concur to define the rating of each player.

E1900: **League ranking mobility affects attendance: A dynamic panel model of European football leagues**
Presenter: **Andras Gyimesi**, University of Pecs, Hungary

In the field of sports economics, there is no strong consensus whether competitive balance of a league increases spectator interest. League ranking mobility is a specific indicator of between-seasons competitive balance, which is defined as the difference between the rankings of two consecutive

time periods. Several measures are adopted and introduced to quantify league ranking mobility. Using an unbalanced panel dataset of 20 European domestic football leagues, it is shown that there is a significant positive relationship between league ranking mobility and average stadium attendance. We argue that a traditional random effect model is not suitable for predicting average attendance, as it cannot handle autocorrelation in the dependent variable. A dynamic panel model, estimated by a GMM estimator, is introduced instead. An innovation of the study is that variables measuring the international success and the average stadium capacity of the league are included as covariates. The effects of other more traditional competitive balance indices are also tested, and found to be non-significant. This research was partially supported by the Human Resource Development Operational Programme, grant No.: HRDOP-3.6.2-16-2017-00003, Cooperative Research Network in Economy of Sport, Recreation and Health.

EC800 Room MAL 153 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS Chair: Stefan Sperlich

E0273: **On a projection estimator for Monte Carlo integration**
Presenter: **Huei-Wen Teng**, National Chiao Tung University, Taiwan

For high-dimensional integration, variance reduction techniques help to improve the efficiency of the standard Monte Carlo simulation, but the reduced variance is difficult to derive theoretically. A novel theoretical perspective is proposed to explain the variance reduction technique by connecting the Monte Carlo estimator with the idea of projection in linear algebra analysis. The mean of a function for a random vector can be seen as a projection of the function on the constant function and the variance is the distance between the above two functions. This framework allows us to propose a new category of estimators for variance reduction with the idea of symmetric group, called the projection estimator. To account for both the amount of variance reduction and computation time, we define the efficiency ratio between two estimators. It is interesting that the commonly known antithetic variates and spherical estimators can be regarded as special cases of projection estimators, and its efficiency ratios can be analysed theoretically. For a projection estimator, the efficiency ratio for a polynomial target function of finite order can be derived exactly. As a result, the proposed framework can be extended to approximate the efficiency ratio of the projection estimator for a general target function. Examples are given both theoretically and numerically.

E1624: **Sample-specific stability selection with effective error control**
Presenter: **Heewon Park**, Hiroshima University, Japan

Recently, sample-specific analysis has drawn a large amount of attention for identifying patient-specific characteristics in the progression of cancer. In order to effectively identify sample-specific molecular mechanisms, we propose a novel sample-specific feature selection method based on the stability selection. Although stability selection provides effective results for variable selection, the method's results are sensitive to the value of the regularization parameter because the method performs feature selection based only on the particular parameter value that maximizes the selection probability. To settle on the issue, we propose robust stability selection and show that our method provides an effective theoretical property (i.e., per-family error rate). We then develop a sample-specific stability selection method based on the kernel-based L1-type regularization and weighted random re-sampling technique. The proposed method estimates the selection probabilities of variables using the sample-specific random lasso and then perform feature selection based on robust stability selection. We observe through the numerical studies that our strategies can effectively perform sample-specific analysis.

E1645: **Efficient simulation of high dimensional Gaussian vectors**
Presenter: **Nabil Kahale**, ESCP Europe, France

A Markov chain Monte Carlo method is described to approximately simulate a centered d-dimensional Gaussian vector X with given covariance matrix. The standard Monte Carlo method is based on the Cholesky decomposition, which takes cubic time and has quadratic storage cost in d. In contrast, the additional storage cost of our algorithm is linear in d. We give a bound on the quadratic Wasserstein distance between the distribution of our sample and the target distribution. Our method can be used to estimate the expectation of $h(X)$, where h is a real-valued function of d variables. Under certain conditions, we show that the mean square error of our method is inversely proportional to its running time. We also prove that, under suitable conditions, the total time needed by our method to obtain a given standardized mean square error is quadratic or nearly quadratic in d. A numerical example is given.

E1777: **Lower quantile estimation for artificially censored Weibull samples**
Presenter: **Jarod Smith**, University of Pretoria, South Africa
Co-authors: JT Ferreira, Andriette Bekker

Quantile estimation is a vital aspect of statistical analyses in a variety of fields. For example, lower quantile estimation is crucial to ensure the safety and reliability of wood-built structures. An intuitive approach would be to consider models that fit the tail of the sample instead of the entire range. Quantiles of interest can be estimated by artificially censoring observations beyond a chosen threshold. The choice of threshold is crucial to ensure efficient and unbiased quantile estimates, and usually the 10th empirical percentile is chosen as the threshold. A bootstrap approach has been previously proposed in order to obtain a better threshold for the censored MLE, however, this approach is computationally expensive. A new threshold selection technique is proposed that makes use of a standardised-weighted adjusted truncated Kolmogorov-Smirnov test (SWAKS-MLE). The SWAKS-MLE outperforms in the bootstrap threshold censored Weibull MLE method, in addition to being vastly less computationally intensive.

E1610: **Identifying differential distributions using the 2-Wasserstein distance, with application to single-cell RNA-sequencing**
Presenter: **Roman Schefzik**, German Cancer Research Center (DKFZ), Germany

A typical statistical issue is to check for differential distributions across two conditions. To address this, the use of the 2-Wasserstein distance is reviewed by putting related scattered literature results into an overarching context being useful in various applications. Specifically, the major causes of differences between distributions can be identified using a decomposition of the 2-Wasserstein distance into location, size and shape deviations. Moreover, different two-sample tests involving the 2-Wasserstein distance are presented: first, a semi-parametric, permutation-based test with a generalized Pareto distribution approximation, and second, a test based on asymptotic theory. Simulations using normal distribution models confirm the validity and usefulness of the findings. In an application, the concepts are specifically adapted to detecting differential gene expression distributions in data from single-cell RNA-sequencing, a recent biological breakthrough technology providing information from multiple individual cells. In particular, the adapted approach tests for differential proportions in zero expression using logistic regression, and for differences in non-zero expression using the semi-parametric 2-Wasserstein distance-based test. The competitiveness of the approach is confirmed in a real-data case study, in which known marker genes and biological patterns can be re-identified, along with additional insights. The methods are implemented in the R package waddR.

EG179 Room MAL 254 CONTRIBUTIONS IN STATISTICAL MODELLING I Chair: David Haziza

E1712: **Mastering the body and tail shape of a distribution**
Presenter: **Matthias Wagener**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The normal distribution and its perturbations have left an immense mark on the statistical literature. Hence, several generalized forms were developed to model different skewness, kurtosis, and body shapes. However, it is not easy to distinguish between changes in the relative body and tail shapes when using these generalizations. What we propose is a neat integration approach which enables the visualization and control of the

body and the tail shape separately. This provides a flexible modelling opportunity with an emphasis on parameter inference and interpretation. Two related models, the two-piece body-tail generalised normal (TPBTGN) and the two-piece tail adjusted normal (TPTAN) are swiftly introduced to demonstrate this potential. This flexible modelling methodology is then demonstrated on heavy and light-tailed data.

E1818: **E-consumers' attribute non-attendance switching behavior: Effect of providing information on attributes**
Presenter: **Michel Meulders**, KU Leuven, Belgium
Co-authors: Leonard Maaya, Martina Vandebroek

Choice experiments are used to investigate how product attributes affect product preference. A choice experiment consists of multiple sets of designed product alternatives that are described by a combination of attribute levels. Respondents are asked to choose the most preferred alternative from each choice set. Based on random utility theory, standard multinomial logit choice models can be used to estimate the utility of attribute levels. Standard choice models assume that respondents examine all attributes in the same fully compensatory manner. However, research has indicated that respondents may ignore part of the attributes if the choice task is too complex. Accounting for such Attribute Non-Attendance (ANA) is important as failure to account for it may lead to biased preference estimates. We develop a dynamic mixture latent Markov model to model the dynamics in attribute non-attendance behavior due to providing information about key-attributes in the course of the choice experiment. The model is illustrated with an application on e-consumers' preferences for webshops. The results indicate that providing information about attributes leads to an increase in the attendance probability of those attributes. Moreover, a dynamic ANA model fits the data better than a model that assumes a change in the preference parameters.

E1923: **Finding the hidden link: Sparse common component analysis**
Presenter: **Katrijn Van Deun**, Tilburg University, Netherlands

Recent technological advances have made it possible to study human behavior by linking novel types of data to more traditional types of psychological data, for example linking psychological questionnaire data with genetic risk scores. Revealing the variables that are linked throughout these traditional and novel types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of depression. Little or no theory is available on the link between such traditional and novel types of data, the latter usually consisting of a huge number of variables. The challenge is to select - in an automated way - those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. To fill the methodological gap, we present an extension of simultaneous component analysis. Constraints are introduced to impose block-structure and to force automated selection of the relevant variables. We will present an efficient procedure that is scalable to the setting of a very large number of variables. Using simulated data and an empirical example, we will showcase the benefits of the proposed method and compare with various competing methods, including sparse PCA and rotation techniques.

E1880: **A new flexible Birnbaum Saunders model**
Presenter: **Inmaculada Barranco-Chamorro**, Universidad de Sevilla, Spain
Co-authors: Guillermo Martinez-Florez, Heleno Bolfarine, Hector W Gomez

The Birnbaum-Saunders (BS) distribution was originally introduced to model the fatigue in lifetime of certain materials. During the last decades, mainly due to its good properties, the use of this model spread out to other fields, such as economics and environmental sciences. In these applied scenarios, quite often, departures of the BS model are found, and therefore it is necessary to introduce some improvements. We focus on those situations in which extra asymmetry or bimodality are present in our data, and a generalization of the BS model should be considered. Our proposal is based on the flexible skew-normal and includes, as particular cases, the BS and skew BS distribution. The key aspects of this model, which properly combined result in the flexible BS model, are asymmetry, bimodality and main features of basic BS model. Theoretical properties are studied. Parameter estimation is considered via an iterative maximum likelihood algorithm. Two real applications, of interest in environmental sciences, are included, which reveal that our proposal can perform better than other compelling models.

E1411: **Constructing distributions via the beta-generating technique**
Presenter: **Seitebaleng Makgai**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi, Jaco Visagie

The beta-generating technique, is a well known mechanism for influencing the tails, the skewness, as well as the goodness of fit of a distribution. Two classes of multivariate beta distributions are proposed and constructed through this nesting technique, where various baseline distributions are considered. General properties such as the cumulative distribution function (cdf), probability density function (pdf), dependence structures and moments are investigated for the developed distributions. The maximum likelihood estimation is used here for parameter estimates, where the usefulness and performance of the distributions are illustrated under various datasets.

EG083 Room MAL 251 CONTRIBUTIONS IN BIG AND HIGH-DIMENSIONAL DATA ANALYSIS Chair: Rosa Lillo

E0361: **StreamMARS: A streaming multivariate adaptive regression splines algorithm**
Presenter: **Inci Batmaz**, Middle East Technical University, Turkey
Co-authors: Niall Adams

Computers and internet have become inevitable parts of our life in the 1990s, and afterwards, bulk of data are started being recorded in digital platforms automatically. To extract meaningful patterns from such data computational methods are developed in data mining and machine learning domains. Multivariate adaptive regression splines (MARS) is one such method successfully applied to off-line static data for prediction. In about last ten years, we face with the big data problem due to the steady increase in the size of the data. Streaming data is a kind of big data collected from sensor networks, production processes, twitter messages etc. Algorithms processing this type of data should consider both memory and time limitations as well as its changing nature with time. We develop a streaming version of a powerful predictive method MARS for estimating model parameters on-line in a temporarily adaptive manner using forgetting factors. Performance of the algorithm developed is tested on simulated data with different dimensions in static, abrupt and smoothly changing environments; as well as on real-life datasets, and also, compared with those of some benchmarking methods such as sliding windows. Results show that StreamMARS is a promising algorithm for predicting streaming big data.

E1692: **Bayesian hierarchical models for uncertainty quantification in high-dimensional landscape problems**
Presenter: **Oluwole Oyebamiji**, Lancaster University, United Kingdom

A Bayesian hierarchical model is tested for achieving dimension reduction in modelling of large dimensional spatial data. This being a step towards emulating a complex integrated model of land-use change. The method uses a combination of Bayesian principal component and Gaussian process based on nearest neighbour approximation. The approach is to first retrieve the low-dimensional underlying patterns from high-dimensional outputs using a Bayesian principal component analysis where the effective dimensionality of the latent space is determined automatically as part of the Bayesian inference procedure. This is followed by the emulation of the resulting low-dimensional data using a composite nearest-neighbour GP based on an assumption of conditional independence. This reduces model complexity and captures different aspects of the socio-economic scenarios. The approach is computationally efficient and improves the accuracy of estimating the parameters as well as incorporating various sources of uncertainty. The method is being applied to a dataset from the IMPRESSIONS Integrated Assessment Platform (IAP2) model, an extension of the CLIMSAVE IAP, which has been widely applied in climate change impact, adaptation and vulnerability assessment for robust policy analysis.

E1800: **Implementation guidance of sparse principal component analysis on different methods and when to use them**
Presenter: **Rosember Guerra-Urzola**, Tilburg University, Netherlands
Co-authors: Katrijn Van Deun, Juan Vera Lizcano, Klaas Sijtsma

Several sparse PCA methods have been introduced for reasons such as interpretability gains and improvement on statistical efficiency in the high-dimensional setting. Existing procedures for sparse PCA are based on different model formulations of PCA in combination with different optimization criteria and numerical techniques to obtain sparseness. The current sparse PCA literature misses clear guidance on the properties and performance of the different methods, often relying on the misconception that the equivalence of the formulations for ordinary PCA also holds in sparse PCA formulations. A guide on the implementation of sparse PCA is offered. First, we discuss several popular sparse PCA methods in terms of the assumed model, the optimization criterion used to impose sparseness, and the algorithmic procedure. Second, using extensive numerical experiments, we assess the performance of each of these methods on performance measures such as Mean Square Error, Percentage of explained variance, and Misidentification rate under several conditions for the population model. Our study highlights that the different sparse PCA methods may yield very different results. We offer some guiding rules in choosing among the different sparse PCA methods that are tailored to the aim of the PCA analysis and the characteristics of the data.

E0880: **Estimating the unemployment rate using big data**
Presenter: **Elisa Jorge-Gonzalez**, Universidad de La Laguna, Spain

In the last decade, the increased availability of open data has awakened interest to develop indicators for many kinds of phenomena, as well as a need for real-time tools that can create economic time series. Google, Amazon or Apple are an example of companies that use the real-time data generated by the customers activities to extract and use knowledge from raw data to make decisions. Estimates and forecasts of the unemployment rate is an important and difficult task for policymakers. Incorporating information from real-time data has been recently shown to improve estimates and short-term forecasts. The main objective is to investigate whether the use of big data can estimate and forecast the unemployment rate in a particular place and moment, thanks to the use and combination of both open data sources and commercial data. Since the methodology is based on open data, this estimation could be applied to any region worldwide.

E1807: **A new approach to penalized quantile regression**
Presenter: **Alvaro Mendez Civieta**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, M Carmen Aguilera-Morillo

Along years, quantile regression has become a key technique used to obtain robust estimators that are able to deal with heteroscedasticity and outliers. In high-dimensional problems, sparsity constrains have shown a great improvement in interpretability and prediction accuracy. One of the best known constraints is the sparse group lasso (SGL). SGL is a penalization technique used in regression problems where the covariates have a natural grouped structure, providing solutions that are both between and within group sparse. We introduce a flexible version, the adaptive sparse group lasso (ASGL), that adds weights to the penalization. Usually, these weights are taken as a function of the original non-penalized model. This approach is only feasible in low-dimensional problems. We propose a solution that allows using adaptive weights in high-dimensional scenarios. We show the benefits of our proposal in a real genetic dataset.

CI022 Room Beveridge Hall ADVANCES IN FINANCIAL ECONOMETRICS Chair: Roxana Halbleib

C0158: **The leverage effect puzzle revisited: Identification in discrete time**
Presenter: **Eric Renault**, University of Warwick, United Kingdom

In spite of evidence that leverage effect (negative correlation between volatility and return) should be present, the estimation of this effect is notoriously difficult. We propose a new discrete time stochastic volatility model with leverage effect that is a natural discrete time analog of popular continuous time affine option pricing model. With an exponentially affine stochastic discount factor, the historical and the risk neutral models belong to the same family of joint probability distributions for return and volatility processes. The discrete time approach allows making more transparent the role of various parameters: leverage versus volatility feedback effect, connection with daily realized volatility, impact of leverage on the volatility smile, etc. Even more importantly it sheds some new light on the identification of leverage effect and of the various risk premium parameters through link functions in closed form. The price of volatility risk is identified from underlying asset return data, even without option price data, if and only if leverage effect is present. However, if leverage effect is close to zero, identification of the volatility risk price may be weak, leading to a new procedure of identification robust inference based on link functions.

C0159: **Analyzing commodity futures using factor state-space models with Wishart stochastic volatility**
Presenter: **Roman Liesenfeld**, University of Cologne, Germany
Co-authors: Tore Selland Kleppe, Guilherme Moura, Atle Oglend

A factor state-space approach with stochastic volatility is proposed to model and forecast the term structure of future contracts on commodities. The approach builds upon the dynamic 3-factor Nelson-Siegel model and its 4-factor Svensson extension and assumes for the latent level, slope and curvature factors a Gaussian vector autoregression with a multivariate Wishart stochastic volatility process. Exploiting the conjugacy of the Wishart and the Gaussian distribution, we develop a computationally fast and easy to implement MCMC algorithm for the Bayesian posterior analysis. An empirical application to daily prices for contracts on crude oil with stipulated delivery dates ranging from one to 24 months ahead show that the estimated 4-factor Svensson model with two curvature factors provides a good parsimonious representation of the serial correlation in the individual prices and their volatility. It also shows that this model has a good out-of-sample forecast performance.

C0237: **Looking forward, looking back: How machine learning predictions compare to expectations from option markets**
Presenter: **Joachim Grammig**, Eberhard Karls Universitaet Tuebingen, Germany
Co-authors: Constantin Hanenberg, Christian Schlag, Jantje Soenksen

There are two opposing strategies to estimate the conditional expectation of a stock return, which is the optimal forecast using mean-squared error (MSE) loss. The first strategy is theory-based, parsimonious, and forward-looking. It utilizes the information contained in option prices, risk-free rates, and the price of the underlying to deliver conditional expected stock returns. The second strategy is theory-free, data science-driven, and backward-looking. It employs machine learning algorithms which are designed to find patterns in historical data to produce MSE-optimal return predictions. Although the two strategies are very different, they both pursue the same goal: Finding the best possible approximation of the conditional expected return. A level playing field is provided to assess the comparative advantages of the forward- and backward-looking approaches towards providing MSE-optimal return predictions. The analysis focuses on the S&P 500 constituents with firm-level data ranging from 1972 to 2017 and volatility surface data ranging from 1996 to 2017. The results of the study suggest a tight head-to-head race of the two strategies since neither of them is able to substantially outperform the other.

CO204 Room Bloomsbury RECENT TRENDS IN COMMODITY MARKETS Chair: Christoph Sulewski

C0927: **Sheep in wolves clothing: Using false signals of demand to execute a market power manipulation**
Presenter: **Craig Pirrong**, University of Houston, United States

The most famous type of market manipulation is a corner, where a trader accumulates a futures position that exceeds the supply of the commodity at delivery points. This type of manipulation has led regulators to compare a long's position to deliverable supply to determine whether the long had manipulated. This test implicitly assumes that shorts believe a long will not consume what is delivered. This raises the question whether there

may be situations in which shorts believe that a long might consume what they deliver. A signaling model is presented in which some longs place a high value on the physical commodity and consume what is delivered to them, but there are other longs who place a low value on it and will not consume it. Shorts do not know which type of long stands for delivery. A low-valuation type exploits this uncertainty to execute a manipulation by misrepresenting his demand for the commodity by offering to sell his positions at a price that equals the valuation of the high-value type. Shorts do not know his true demand, and assign some positive probability to the possibility that the long's offer price reflects his actual valuation and that he will consume the deliveries. There is a pooling equilibrium in which low-value demanders mimic high-value demanders. This allows them to liquidate some of their futures positions at an artificially high price and profit even though they lose money when they resell what shorts deliver to them.

C1438: **Speculative pressure**
Presenter: **Joelle Miffre**, Audencia Business School, France
Co-authors: John Fan, Adrian Fernandez-Perez, Ana-Maria Fuertes

The pricing content of speculative pressure in diverse futures classes is investigated. Long-short portfolios of futures contracts sorted by speculative pressure capture a significant premium in commodity, currency and equity markets but not in fixed income markets. Exposure to commodity, currency and equity index futures speculative pressure is priced in the broad cross-section after controlling for momentum, carry, global liquidity and volatility risks. The findings are confirmed by robustness tests using alternative speculative pressure signals, portfolio construction techniques and subsamples inter alia. We argue that there is an efficient hedgers-speculators risk transfer in commodity, currency and equity index futures markets.

C1523: **Gold volatility and the safe haven effect**
Presenter: **Konstantin Kuck**, University of Hohenheim, Germany

The relationship between gold and the stock market has attracted considerable attention in both the academic literature and financial media since gold is perceived a 'safe haven' for equity. This describes the empirical phenomenon that gold holds its value or exhibits positive returns in a situation of extreme stock price declines. Whilst studies investigating the role of gold as safe haven typically focus on returns, we also investigate the relationship between gold volatility and stock returns. In particular, we are interested in the behaviour of gold volatility on days with negative shocks in the equity market. The conditional volatility of gold is of direct relevance for an assessment of the effectiveness of the safe haven property from a portfolio perspective. We find that gold does not move in tandem with stocks but has a significantly higher volatility in response to negative shocks in the equity market. In essence, our findings are in line the notion of gold as a (weak) safe haven. However, they also imply that the higher gold volatility in response to a negative stock market shock contributes to the risk of a portfolio composed of the two assets.

C1529: **Cocoa market control 2020**
Presenter: **Christopher Gilbert**, Johns Hopkins University, Italy

Cocoa is priced against the ICE cocoa contracts traded in London and New York. Ghana and Cote dIvoire account for slightly in excess of 60% of world cocoa production. The cocoa price remains low, mainly as the consequence of increases in Ivorian production. The two governments have announced a plan to boost cocoa revenues by requiring exporters to pay a minimum differential of $400/ton (around 22%) over the London futures price. There are major issues as to whether and how this requirement can be enforced. We look at the consequences of the scheme if it can be successfully implemented. If effective, this scheme would drive a wedge between transactions and exchange prices. The result would be similar to the recent wedge between aluminum exchange and transactions prices resulting from loadout delays from exchange warehouses. We consider the implications for cocoa futures prices, for the differentials obtained other origins, at likely impacts on hedge effectiveness and at the possible impact on Ghanaian and Ivorian farmers.

C1140: **An introduction to ESMA's commitments of traders reports: Do hedgers really hedge?**
Presenter: **Martin Stefan**, University of Muenster, Germany
Co-authors: Claudia Wellenreuther, Martin T Bohl

A novel type of commodity futures positions report issued by the European Securities and Market Authority (ESMA) is introduced. This report is interesting to researchers for two reasons: First, it allows analyzing European commodity markets, which, compared to US-American markets, have hitherto largely been ignored by the literature. Second, this new type of report offers the advantage of breaking down positions not only by the different types of traders but also by the underlying trading motives. These new data for different energy and metal futures contracts are studied. The results suggest that the extent of speculative positions might have been underestimated in earlier studies.

CO420 Room G11 EMPIRICAL APPLICATIONS IN ECONOMICS AND FINANCE Chair: Jesus Otero

C1385: **Forecast revisions in a changing economic environment**
Presenter: **Ana Maria Iregui Bohorquez**, Banco de la Republica, Colombia
Co-authors: Jesus Otero, Hector Nunez

The forecasts of inflation and exchange rate contained in the Monthly Survey of Economic Analysists of the Banco de la Republica (Central Bank of Colombia) for the period 2004-2018 are analyzed. The forecasts of these two variables of interest, which are made for December of the current year as well as December of the following year, comprise an unbalanced panel of data in three dimensions with multiple individual forecasters, target years, and forecast horizons. The fixed-event nature of the forecasts enables us to examine their (weak and strong) efficiency by looking at the way they are revised by the analysts from survey to survey. According to our results, changing economic conditions appear to affect how efficiently analysts include new information when updating their inflation and exchange rate forecasts.

C0478: **Banking competition, institutional investors and financial constraints: Evidence from Europe**
Presenter: **Carlos Pombo**, Universidad de los Andes, Colombia
Co-authors: Mauricio Jara-Bertin

Using a sample of listed firms that belongs to 21 European Countries for the period 2000-2015, the aim is to analyze the relationship between banking concentration, institutional investors and their impact on financial constraints. The results show that according to the information hypothesis, bank concentration reduces financial constraints, measured as the sensitivity of the investment to cash flows. However, this effect is mitigated to the extent that institutional investors have a greater participation in the ownership of the company, which is consistent with a potential resource competition for reducing agency costs. The results are robust by analyzing different sources of heterogeneity. We include the heterogeneity at the institutional investor type i.e., grey versus independent, firm size and opacity. Our findings suggest that the investors of the independent type are the ones that would diminish the financial restrictions of the firms, and that the smaller, less opaque companies will replicate the results found in the base scenario.

C0195: **Exchange rate dependencies using a copula approach**
Presenter: **Luis Fernando Melo Velandia**, Banco de la Republica, Colombia
Co-authors: Juan Cubillos, Jose Eduardo Gomez

Exchange rate dependencies between seven countries from four different regions of the world are studied. The sample includes two developed countries, the United Kingdom and Germany, two large emerging Asian economies, South Korea and Indonesia, two Latin American countries, Brazil and Chile, and South Africa. The currencies of all of these countries are actively traded in global forex markets and all of them are important

for large international portfolio composition and rebalancing. We construct multivariate copula functions using a regular vine copula approach, allowing for very flexible dependency structures. We find evidence of exchange rate contagion for our set of countries. However, important asymmetries are worth noting. First, contagion occurs only during periods of exchange rate appreciation of the different currencies with respect to the United States Dollar. We do not find evidence of contagion for any pair of exchange rates during periods of currency depreciations. Second, contagion is more frequent in pairs of countries that include either the United Kingdom or Germany. Third, contagion occurs more within countries of a same region, for instance, between Brazil and Chile, and between Korea and Indonesia. This result shows that, in episodes of large currency appreciation, hedging strategies for global investors taking positions in large markets require regional diversification.

C1019: **Structural factor analysis of interest rate pass through in four large Euro area economies**
Presenter: **Anindya Banerjee**, University of Birmingham, United Kingdom
Co-authors: Victor Bystrov, Paul Mizen

The influence of monetary policy decisions of the ECB on mortgage and business lending rates offered by banks in the four major euro area countries (Germany, France, Italy and Spain) is examined. Since there are many different policy measures that have been undertaken, we utilize a dynamic factor model, which allows examination of impulse responses to policy shocks conditioned upon structurally identified latent factors. The distinct feature is that it explores the effects of three policy transmission lines - short-term rates, long-term rates and perceived risk - ultimately directed towards bank lending rates. The analysis of the pass through is carried out in pre-crisis and post-crisis sub-samples to demonstrate the changing influence of different policy measures onlending rates.

C0203: **Rigidities and adjustments of daily prices to costs: Evidence from supermarket data**
Presenter: **Jesus Otero**, Universidad del Rosario, Colombia
Co-authors: Monica Giulietti, Michael Waterson

The extent of inertia in grocery retail prices is assessed by using data on prices and costs from a large supermarket chain in Colombia. Relative to previous work, we benefit from the daily frequency of the data, the availability of reliable replacement cost data and the wide variety of products. The data supports the existence of significant nominal rigidities in reference prices (three months) and even more so in reference costs (about five months). The price and cost rigidities differ depending on the type of product, being on average smaller in the case of perishable goods. Using an Error Correction Model, we examine the path of prices relative to costs, to determine the speed of adjustment of prices to shocks, finding significant mean reversion but also uncovering asymmetries in response to upward and downward movements in margins in many cases.

CO382 Room G21A ASSET PRICING AND RISK EXPOSURES IN CRYPTOCURRENCY MARKETS Chair: Massimo Guidolin

C1392: **Time-varying risk exposures of cryptocurrencies: Are they a new asset class?**
Presenter: **Massimo Guidolin**, Bocconi University, Italy
Co-authors: Daniele Bianchi, Manuela Pedio

Despite the interest in cryptocurrencies has recently increased, in the literature a consensus has not yet been reached on whether they represent or not a new asset class, spanning risks and payoffs sufficiently different from the traditional ones. We contribute to this debate by studying the exposure of cryptocurrency returns to stock market risk factors (namely, the six Fama French factors), to precious metal commodity returns, and to cryptocurrency-specific risk-factors (namely, crypto-momentum, a sentiment index based on Google searches, and supply factors, i.e., electricity and computer power). Because economic facts lead us to believe that those exposures are likely to be time-varying, we rely on Bayesian methods, which incorporate dynamic model averaging. These methods not only feature time-varying coefficients, but also allow for the entire forecasting model to change over time. We estimate our flexible models on weekly data for four popular cryptocurrencies, namely Bitcoin, Ethereum, Litecoin and Ripple. We find that cryptocurrencies are not systematically exposed to stock market factors, precious metal commodities or supply factors with exception to some occasional spikes of the coefficients during our sample. On the contrary, they display a time-varying but significant exposure to a sentiment index and to crypto-momentum.

C1395: **The cross-section of cryptocurrency returns**
Presenter: **Kirill Shakhnov**, University Of Surrey, United Kingdom
Co-authors: Nicola Borri

Investors can invest in cryptocurrencies on a multitude of exchanges located in different countries and against different fiat currencies and cryptocurrencies. We start off large and persistent differences in bitcoin prices, converted to U.S. dollars, across exchange-currency pairs. These differences provide information about each pair's exposure to common risk factors, which we identify by building portfolios sorted on bitcoin past price deviations and accounting for transaction costs and execution risks. A single crypto factor, Carry, explains most of the variation in bitcoin excess returns. Portfolios with higher average returns are riskier because have higher (lower) returns in good (bad) times for cryptocurrency investors: when trading volume is larger (smaller) and the probability of exchange shutdowns lower (higher).

C1400: **Trading volume in cryptocurrency markets**
Presenter: **Alexander Dickerson**, Warwick Business School - University of Warwick, United Kingdom
Co-authors: Daniele Bianchi

The value of trading volume in cryptocurrency markets is studied, contributing to a growing literature that aims to understand the role of cryptocurrencies as investment. The main results show that the interaction between lagged volume and past returns have a significant predicting power for future returns. Such predictive power is economically significant; an investment strategy that conditions on past returns and volume generates a substantial Sharpe ratio with zero correlation with Bitcoin and Ethereum dollar returns. These results are consistent with existing theoretical models that postulate that is primarily "speculation" on private information that generates the observed returns dynamics.

C1401: **Return spillovers and connectedness in the cryptocurrency market**
Presenter: **Engin Iyidogan**, SKEMA Business School, France
Co-authors: Kamil Yilmaz

The data show that system-wide connectedness in the cryptocurrency market, commonly observed as the level of systemic risk in the literature, changes over time with respect to the adoption rate of cryptocurrencies. Cryptocurrency protocol disputes, malicious activities in the cryptocurrency environment, and negative financial regulations on cryptocurrencies are the major sources of increase in connectedness for different periods. Instead, the pairwise connectedness between cryptocurrency pairs can be explained by cryptocurrency fundamentals.

C1657: **Price endogeneity and volatility in the cryptocurrency market**
Presenter: **Jan Sila**, Univerzita Karlova, Czech Republic
Co-authors: Michael Mark

The aim is to describe the endogenous dynamics of the emerging cryptocurrency markets by constructing so-called "reflexivity index". A univariate self-exciting Hawkes process with various kernels is fit to high-frequency Bitcoin transactions from BitMEX exchange. We then explore the predictive power of the reflexivity index for volatility forecasting. Since cryptocurrencies are considered as highly volatile instruments, we enrich the exponential GARCH model with external regressor of market activity, to markedly improve the volatility forecasts particularly for the one-step-ahead frequency. This introduces a novel model-stacking concept, where we improve traditional volatility forecasting methods with features from a slightly different approach to financial modelling.

CO196 Room G5 ECONOMETRICS METHODS AND MODELS FOR HIGH DIMENSIONAL DATA ANALYSIS Chair: Camilla Mastromarco

C0826: **Whether commodity price volatility spills over onto African stock markets**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Giampiero Gallo, Vincenzo Candila, Marinella Boccia

The aim is to examine the volatility transmission from energy and metal commodities to six major African exporters' stock markets (Egypt for oil and gold, Nigeria for oil and gas, South Africa for coal and gold, Tunisia for oil, Uganda for gold and Zambia for copper). Modelling commodity volatility with the Double Asymmetric GARCH-MIDAS model with a Student's t-distribution allows to detect the presence of impact and inertial stock market volatility spillovers at different lags and to take into account the leptokurtosis of the commodity series. We then derive the profile of Volatility Impulse Responses of the stock markets to commodity shocks.

C0822: **Daily, intraday and overnight betas**
Presenter: **Alessandra Insana**, University of Messina, Italy

There are substantial differences between daily intraday and overnight betas. Dividing the total daily return in intraday and overnight return, we evaluate our three betas by two models. We start considering the classical Capital Asset Pricing Model (CAPM), assuming a constant systematic risk, and so a constant and unconditional beta over time. Then, we apply a nonparametric method for time varying conditional betas. For the analysis, we use US stocks traded on the NYSE, AMEX, and NASDAQ markets, computing our three betas considering single stocks and aggregating them in portfolios by market capitalization. As benchmark for the market index, we estimate a specific value weighted index related to each period.

C1353: **Endogenous spatial technological clubs across Europe: A 3D panel data model with cross sectional dependence**
Presenter: **Camilla Mastromarco**, University of Salento - Lecce, Italy

By proposing a new approach which merges previous methodologies, both strong and weak spatial dependence is considered, and European regions spatial clusters endogenously are modelled in a three-dimension panel data model. The aim is to study whether the European regions form regional clusters that differ from the political borders. We analyse interdependence across regions of production technology clusters and estimate the degree of regional technological interdependence generated by the level of spatial externalities. The net effect of these spatial externalities on the level of productivity of each region depends on the relative connectivity between this region and its neighbors. The more a given region is connected to its neighbors, the more it benefits from spatial externalities. The more efficient is a region, the more it profits from technological externalities. Further, we investigate the regional patterns of those clusters over time. We use data from Cambridge Econometrics European Regional Database contains annual observations for the period 1980-2015 for NUTS3 EU 27 regions and Norway on: employment (thousands of people); hours worked; gross fixed capital formation in millions of Euros 2005 prices; gross value added in millions of Euros 2005 prices.

C1682: **Sample separation and the sensitivity of investment to cash flow: On the monotonicity condition**
Presenter: **Leone Leonida**, King's College London, United Kingdom
Co-authors: Alfonsina Iona

The purpose is to study whether the monotonicity condition of the investment-cash flow sensitivity is satisfied empirically. We show that if this condition holds, then the point of sample separation does not affect the monotonic relationship between the sensitivities of any two complementary classes of observations. Our test, based upon observable averages of the investment-cash flow sensitivity, rejects the monotonicity condition for any common metric of financing constraints we use. The testing procedure we propose reconciles the conflicting findings of the literature about the shape of the investment-cash flow sensitivity.

C1684: **Economic freedom, corporate investment and financing constraints.**
Presenter: **Alfonsina Iona**, Queen Mary University of London, United Kingdom
Co-authors: Leone Leonida

During the 2008 financial crisis, economic freedom and firm investment begin to undertake a declining path, while financial constraints appear to become more severe for the US firms. Using a standard model of investment, we study the relationship between economic freedom, corporate investment and financial constraints in a less than perfectly competitive market and in presence of imperfect capital markets. We test the theoretical predictions of our model by using a large panel of nonfinancial US firms. Our results show that economic freedom positively affects corporate investment, and it is able to mitigate the negative impact of financial constraints on investment. Moreover, we find that these effects act mainly through credit market freedom, efficiency of legal system and access to money.

CO442 Room Gordon ROBUSTNESS TO SHOCKS AND DEPENDENCE IN NETWORKS AND FINANCIAL DATA Chair: Artem Prokhorov

C1118: **Identifying important nodes in input-output networks**
Presenter: **Alexander Semenov**, University of Jyvaskyla, Finland

The economic system may be represented as a complex network, consisting of different sectors linked by input-output relations. In such a network, shocks occurring in one of the sectors may result in wide disruptions of other sectors' production. Previous research in network science has established multiple ways to measure the importance of a node or an edge in a network, such as degree centrality, closeness centrality, PageRank, or betweenness centrality. Shocks occurring in the most important sectors may lead to more pronounced impacts. Several attempts have been made to analyze the topology of the networks constructed from input-output tables; however, there has been little analysis of the effect of disruptions in the important sectors of the other parts of the network. We construct networks from published data on input-output relations, and study how shocks occurring in the most important sectors of the economy, quantified according to network science-based measures, affect production of other sectors. Next, we propose an optimization problem aimed at finding the most influential sectors and explore how effects of important nodes differ over networks constructed from different economies.

C1291: **Machine learning algorithms for profiling of Russian digital entrepreneurs: Social network data analysis**
Presenter: **Margarita Gladkova**, Saint Petersburg State University, Russia
Co-authors: Evgenii Gilenko

When implementing their business ideas, young entrepreneurs are increasingly turning to the digital form of business, because it allows them to quickly attract the target audience (specifically in Russia, more than 75% of the population are Internet users, 83% of whom have social network accounts), quickly respond to changing audience needs, get more information about customers and the market as a whole. The biggest Russian social network VKontakte accounts more than 400 thousand businesses. The aim is to study digital business presented in social networks, entrepreneurs' motives for its creation, features of the digital business and its interaction with the clients. The detailed portraits of entrepreneurs operating on a digital platform and classification of businesses are created with the help of the modern methods of statistical analysis and machine learning.

C1318: **Sign tests for dependent observations**
Presenter: **Rustam Ibragimov**, Imperial College London and St. Petersburg State University, United Kingdom
Co-authors: Donald Brown

New sign tests for testing equality of conditional distributions of two (arbitrary) adapted processes as well as for testing conditionally symmetric

martingale-difference assumptions are introduced. The analysis is based on results that demonstrate that randomization over ties in sign tests for equality of conditional distributions of two adapted sequences produces a stream of i.i.d. symmetric Bernoulli random variables. This reduces the problem of evaluating the critical values of the tests to computing the quantiles or moments of Binomial or normal distributions. Similar properties also hold under randomization over zero values of signs of a conditionally symmetric martingale-difference sequence.

C1527: **Lasso and Dantzig selector for Bernstein copula**
Presenter: **Artem Prokhorov**, University of Sydney, Australia

The LASSO and Dantzig selector are traditionally used for point estimation and model selection, especially within least squares problems where the number of parameters may exceed the number of observations. These methods are proposed to be used to regularise Bernstein copula based MLE problems. Besides offering a computationally efficient penalised estimator of the Bernstein copula, we consider estimation of the parameters in a correctly specified marginal distribution using the information in the copula. We show that the sparsity imposed by the regularisations is innocuous with respect to the non-asymptotic behavior of the sieve MLE, while it permits a substantial increase in computational efficiency compared to the unrestricted sieve MLE. We also study the parameter path behavior over a feasible range of tolerance levels and consider a version of the double Dantzig selector which resolves the arbitrariness in choosing the tolerance.

C1265: **Robust estimation of stochastic frontier models**
Presenter: **Evgenii Gilenko**, Saint Petersburg State University, Russia
Co-authors: Margarita Gladkova, Artem Prokhorov, Rustam Ibragimov, Glafira Kukhareva

Robust methodological tools are developed for stochastic frontier analysis which will permit major improvement of production efficiency by means of a reliable estimation of production functions. A key outcome is the creation of a new generation of stochastic productivity models that can accommodate such important features of the real world as flexible specification of production function, endogeneity of inputs and dependence between inputs and environmental variables. The aim is to contribute to the development of quantitative tools capable of better representing the complex reality of modern production processes. The application of these methods is illustrated on actual data.

CO198 Room Montague TERM STRUCTURE OF INTEREST RATES Chair: Andrea Berardi

C0790: **Yield curve forecasting in the euro area**
Presenter: **Marco Taboga**, Banca d'Italia, Italy
Co-authors: Francesco Corsello

The purpose is to assess the ability of competing models and estimation methodologies to provide accurate point and density forecasts of government yield curves. We focus on euro area countries, which have seldom been analyzed by the literature despite the importance of yield curve forecasts in the Eurosystem policy framework. We find that among several models, including some that have been successful in forecasting the US yield curve, no one is able to consistently produce point forecasts that are superior to random walk forecasts. We instead find that density forecasts produced by Bayesian stochastic volatility models with a random walk component are correctly calibrated and provide the best performance in terms of continuous ranked probability scores.

C1550: **A shadow rate without a lower bound constraint**
Presenter: **Rafael Barros De Rezende**, Bank of England, United Kingdom
Co-authors: Annukka Ristiniemi

A shadow rate is proposed that measures the interest rate effects of monetary policy when the lower bound is not binding. Using daily yield curve data we estimate shadow rates for the US, Sweden, the euro area and the UK, and document that they fall (rise) as monetary policy becomes more expansionary (contractionary). This ability of the shadow rate to track the stance of monetary policy is identified on announcements of policy rate cuts (hikes), balance sheet expansions (contractions) and forward guidance, with shadow rates responding timely, and in line with government bond yields. We show two applications for our shadow rate. First, we decompose shadow rate responses to monetary policy announcements into conventional and unconventional monetary policy surprises, and assess the pass-through of each type of policy to exchange rates. We find that exchange rates respond more to conventional than to unconventional monetary policy. Lastly, a counterfactual experiment in a DSGE model suggests that inflation in Sweden would have been around half a percentage point lower had the Riksbank not used unconventional monetary policy since February 2015.

C1576: **Natural rate chimera and bond pricing reality**
Presenter: **Gavin Goy**, De Nederlandsche Bank, Netherlands
Co-authors: Claus Brand, Wolfgang Lemke

Taking into account secular macroeconomic trends, the decline in yields observed since the 1980s appears more due to a fall in equilibrium interest rates and less to a decline in term premia than typically reported. We incorporate an arbitrage-free term-structure model into a small semi-structural macro-model to jointly estimate potential output growth, output gaps, core inflation, real equilibrium interest rates, and term premia (for the US and the euro area). We illustrate that exploiting cross-sectional information in yields and closing the original macro-model with a short-rate equation increases the precision of natural rate estimates. We use a Bayesian approach to estimate all model components simultaneously. Taking into account the secular fall in equilibrium rates, term premia exhibit cyclical behavior over the business cycle, rather than a trend decline.

C1689: **Macro-yields modelling in the presence of asymmetrically distributed interest rates**
Presenter: **Jingwen Shi**, University of Warwick, United Kingdom

The potential of the copula framework to handle non-Gaussianity in the context of macro-finance term structure modelling and forecasting is studied. The proposed non-Gaussian macro-yields model accounts for the asymmetry and tailedness in yield distributions through non-parametric marginal densities, as well as explicitly addressing the cross-sectional and serial dependence via the state-space inversion copula, and in doing so, retains a latent dynamic factor structure amenable to efficient implementation. Regardless of the maturities and forecast horizons, exploiting the informational content of macroeconomic data in a non-Gaussian setting improves both in-sample and out-of-sample forecasting performance relative to the Gaussian macro-yields model over the 1970:M12016:M12 period. Furthermore, the non-Gaussian macro-yields model demonstrates overwhelming superiority in predicting excess bond returns over the prominent macro-financial predictors and the expectations hypothesis. It also compares favourably with the random walk in forecasting the yield curve over medium- to long-term horizons. Lastly, this copula-based approach affords a technically convenient means of accommodating high-dimensional macroeconomic datasets.

C0874: **Why stochastic volatility matters for the behaviour of long-term interest rates**
Presenter: **Andrea Berardi**, University of Venice, Italy
Co-authors: Stephen Schaefer

The long end of the forward rate curve is consistently and significantly downward sloping. We show that this arises from the volatility of long-term yields and the higher convexity of long-term bonds. The downward slope of long-term forward rates therefore provides a window on the impact of volatility on the term structure. However, both volatility and the size of the downward slope vary significantly over time and a term structure model with stochastic volatility is necessary to account for this feature of the data. We decompose the yield curve into expectations of future rates, term premia and a convexity term and show that the downward slope is almost entirely determined by term premia and convexity effects. We also show that, while the downward slope varies strongly with volatility, its average size is smaller than would be predicted if risk premia were zero,

thus confirming the importance of risk premia in determining the yield curve, especially at the long end.

CO394 Room Woburn MACROECONOMIC POLICIES AND MACROECONOMETRICS Chair: Etsuro Shioji

C0860: **On posterior inferences of misspecified DSGE models: The minimal econometric interpretation**
Presenter: **Takashi Kano**, Hitotsubashi University, Japan

The minimal econometric interpretation (MEI) has been previously developed for formal model evaluation and comparison among misspecified DSGE models in accompany by a theoretical reference model. The MEI approach recognizes a DSGE model as an incomplete econometric tool that provides only prior distributions of targeted population moments, but has no implication for actual data and sample moments. A reference model generates posterior distributions of targeted population moments conditional on data. The MEI approach, however, is absent from posterior updating of structural parameters of DSGE models. The conventional MEI approach is extended by proposing a simple method of drawing posterior inferences of DSGE models. The method exploits a Dirichlet-multinomial model to provide a mixture of prior and posterior distributions of targeted population moments implied respectively by DSGE and reference models. This simulation-based population moment-matching method is applied to draw posterior inferences of nonlinear structural asset pricing models.

C0921: **Pass-through of oil supply shocks to domestic gasoline prices: Evidence from daily data**
Presenter: **Etsuro Shioji**, Hitotsubashi, Japan

Oil prices react to various types of shocks, and their impacts could reach our lives quickly. For example, on one day, we might hear on the radio that OPEC has decided to cut their oil production for the next year, and, only a few days later, find out that our local gasoline station has just raised the price. The aim is to examine daily data on gasoline prices, produced by a price comparison site in Japan, to estimate how they respond to a shock that hit the world oil market. In doing so, we take seriously the possibility that an increase in oil prices might cause different reactions depending on the source of the change. The focus is on one particular type of shock, namely, changes in expectations about future supplies of crude oil. Identification is achieved via estimating a version of the Structural VAR with External Instruments (SVAR-IV or proxy-VAR) coupled with High Frequency Identification (HFI). The result confirms that pass-through is indeed very fast: about 70 percent of the entire adjustment process is completed within just 18 business days.

C1021: **Forecasting public investment using daily stock returns**
Presenter: **Hiroshi Morita**, Hosei University, Japan

Financial market variables contain a lot of information to forecast the variations of macroeconomy. By taking advantage of such a desirable property of financial data, the predictability of the Japanese public investment is investigated by using daily excess stock returns of construction industry to contribute to the recent discussion on fiscal foresight. To examine the relationship between monthly public investment and daily stock returns without any time aggregation, we employ the VAR model with MIDAS regression, in which the optimal weights for connecting high frequency data and low frequency data are estimated in addition to the VAR coefficients and variance-covariance structure. Our results reveal that the VAR model with MIDAS regression reduces the mean square prediction error (MSPE) in the out-of-sample forecast by as much as 14 percent in comparison with the no-change forecast. Moreover, based on the local projection method, we also find that fiscal news shock estimated in our VAR model has a delayed positive effect on output after significant negative effect for almost the first one year.

C1423: **Measuring cross country linkages with a panel unobserved component model**
Presenter: **Gianni Amisano**, Federal Reserve Board, United States

The purpose is to use a panel unobserved component model for output and unemployment to study the interconnectedness among several countries, along different dimensions. In particular, our model allows for commonalities in cyclical and long term movements. Using a time varying parameter version of the model, we are also able to analyze the evolution of cross country linkages. A further extension of the model allows us to examine whether commonalities can be modeled with correlation across idiosyncratic components. The model and its extensions are given a network interpretation.

C1474: **A global look into corporate cash after the global financial crisis**
Presenter: **Kei Ichiro Inaba**, Bank of Japan, Japan

The purpose is to investigate the developments and determinants of cash holdings by publicly traded firms for 20 advanced and emerging countries over the last decade. Ratios of the firms' aggregate cash to their total assets were on upward trends in the majority of the sample countries. Panel-data regressions find that higher cash ratios were associated with fewer non-cash current assets, smaller cost of carry, larger contemporaneous cash inflows, fewer interest-bearing liabilities, greater expected investment opportunities, including R&D projects, larger uncertainty, and the state of corporate governance. The last one is related to the agency motive and, to be specific, higher cash ratios were associated with managers with worse ethicality, lower accountability to investors and board members, weaker investor protection, harsher auditing and reporting standards, and greater potential to face holdup behaviors taken by lending banks. The agency motive was greater than the precautionary and transaction-costs motives in terms of standardized impact while being marginal in terms of explanatory power over total variations in the cash ratios.

CO675 Room Chancellor's Hall DYNAMIC FACTOR MODELS AND LARGE-SCALE APPLICATIONS Chair: Matthias Fengler

C0958: **Large-scale dynamic predictive regressions**
Presenter: **Daniele Bianchi**, Queen Mary University of London, United Kingdom
Co-authors: Kenichiro McAlinn

A large-scale dynamic predictive strategy for forecasting an economic decision making in a data-rich environment is proposed and evaluated. Under this framework, clusters of predictors generate different predictive densities that are later synthesized within an implied time-varying latent factor model. We test our procedure by predicting both the inflation rate and the equity premium across different industries in the U.S., based on a large set of macroeconomic and financial variables. The main results show that our framework generates both statistically and economically significant out-of-sample outperformance compared to a variety of sparse and dense regression-based models while maintaining critical economic interpretability.

C0577: **Contagion and latent factors in large systems**
Presenter: **Federico Carlini**, USI, Lugano, Switzerland
Co-authors: Patrick Gagliardini

A large-dimensional time series model is considered that disentangles dependence patterns due to either the effect of latent common stochastic factors or direct causality effects (i.e. contagion). Our model is $Y_t = CY_{t-1} + Bf_t + u_t$, $t = 1,\dots,T$ where Y_t has dimension N, f_t is a K-dimensional latent factor independent of u_t, a weakly dependent process over time, and C is a contagion matrix. We study identification and estimation when N and T are large. The contagion matrix C has N^2 parameters and we impose the structure $C = (1/N)(\alpha\beta' + \varepsilon)$ for identifiability and interpretability purposes, where α and β are $N \times r$ factors independent by ε, a $N \times N$ random matrix with i.i.d. elements across. We prove that $Y_t = \alpha g_t + Bf_t + u_t = \Lambda h_t + u_t$, where g_t is a $r-$dimensional unobservable factor that captures contagion, $h_t = (g_t', f_t')'$ and $\Lambda = (\alpha : B)$. We estimate Λ and h_t with PCA. We identify r and K through tests on the canonical correlation of h_t and h_{t-1}. Moreover, we estimate g_t, f_t, α and β. Finally, we find the asymptotic properties of the estimators and we illustrate the estimation method with an empirical application.

C0368: **Score-driven exponential random graphs: A new class of time-varying parameter models for dynamical networks**
Presenter: **Giacomo Bormetti**, University of Bologna, Italy
Co-authors: Domenico Di Gangi, Fabrizio Lillo

Motivated by the evidence that real-world networks evolve in time and may exhibit non-stationary features, an extension of the Exponential Random Graph Models (ERGMs) accommodating the time variation of network parameters is proposed. Within the ERGM framework, a network realization is sampled from a static probability distribution defined parametrically in terms of network statistics. Inspired by the fast growing literature on Dynamic Conditional Score-driven models, each parameter evolves according to an updating rule driven by the score of the conditional distribution. We demonstrate the flexibility of the score-driven ERGMs, both as data generating processes and as filters, and we prove the advantages of the dynamic version with respect to the static one. The proposed method captures dynamical network dependencies, that emerge from the data, and allows for a test discriminating between static or time-varying parameters. Finally, we corroborate our findings with the application to networks from real financial and political systems exhibiting non-stationary dynamics.

C0346: **Extracting statistical factors when betas are time-varying**
Presenter: **Hao Ma**, USI Lugano and SFI, Switzerland
Co-authors: Patrick Gagliardini

The focus is on identification and inference on the unobservable conditional factor space and its dimension in large unbalanced panels of asset returns. The model specification is essentially nonparametric regarding the way the loadings vary in time as functions of common shocks and individual characteristics. The number of active factors can also be time-varying as an effect of the changing macroeconomic environment. The method relies on recent proposals deploying instrumental variables in large panels with unobservable factors. It accommodates for a large dimension of the vector generating the conditioning information set by machine learning techniques. In an empirical application we infer the conditional factor space in the panel of monthly returns of individual stocks in the CRSP dataset between February 1971 and December 2017.

C0578: **Global estimation of realized spot volatility inthe presence of price jumps**
Presenter: **Matthias Fengler**, University of Sankt Gallen, Switzerland
Co-authors: Wale Dare

A non-parametric procedure is proposed for estimating the realized spot volatility of a price process described by an Ito semimartingale with Levy jumps. The procedure integrates threshold jump elimination techniques with a Gabor frame expansion of the realized trajectory of spot volatility. We show that the procedure converges in probability in $L_2([0,T])$ for a wide class of spot volatility processes, including those with discontinuous paths. The analysis assumes that the time interval between price observations tends to zero. The intended application is spot volatility estimation from high frequency financial data.

CC819 Room G3 CONTRIBUTIONS IN FORECASTING Chair: Uwe Hassler

C1454: **Nonparametric tests for superior predictive ability**
Presenter: **Stelios Arvanitis**, RC-AUEB, Greece
Co-authors: Thierry Post, Valerio Poti, Selcuk Karabati

A nonparametric method for comparing multiple forecast models is developed and implemented. The hypothesis of Optimal Predictive Ability generalizes the Superior Predictive Ability hypothesis from a single given loss function to an entire class of loss functions. Distinction is drawn between General Loss functions, Convex Loss functions and Symmetric Convex Loss functions. The Optimal Predictive Ability hypothesis is formulated in terms of moment inequality conditions. The empirical moment conditions are reduced to an exact and finite system of linear inequalities based on piecewise-linear loss functions. The hypothesis can be tested in a statistically consistent way using a blockwise Empirical Likelihood Ratio test statistic. A computationally feasible test procedure computes the test statistic using Convex Optimization methods and estimates conservative, data-dependent critical values using a majorizing chi-square limit distribution and a moment selection method. An empirical application to inflation forecasting reveals that a very large majority of thousands of forecast models are redundant, leaving predominantly Phillips Curve type models, when convexity and symmetry are assumed.

C1732: **Fast solution to the automatic identification of unobserved components models**
Presenter: **Diego Pedregal**, University of Castilla-La Mancha, Spain
Co-authors: Juan R Trapero

A fast algorithm and software for the automatic identification of Unobserved Components models are presented. Solutions of this sort is compulsory nowadays once the big data era has come and is staying among us. Crafted identification procedures are reserved for problems in pure scientific contexts where the number of time series are manageable, but are not practical for organizations that want to process a tsunami of information efficiently, online and in record time. Automatic identification tools are the usual way to deal with modeling problems in many contexts, typically Machine Learning and some statistical areas, but it has never been tried out in Unobserved Components models (UC). A new piece of software is introduced with is developed in R with the core written in C++ with the help of RcppArmadillo for the automatic identification and forecasting of UCs with some enhancing features. The package searches among many combinations of different specifications for each of the components according to a stepwise algorithm to reduce the universe of possible models and selects the one with the best metrics. The forecasting results suggest that UC models are powerful potential forecasting competitors to other well-known methods both in computation time and forecasting accuracy. Though there are several pieces of software available for UC modeling, this is the first implementation of an automatic algorithm for this class of models, to the authors' knowledge.

C1738: **Reversed order monitoring CUSUM test with factor structure**
Presenter: **Shou-Yung Yin**, National Taipei University, Taiwan
Co-authors: Chang-Ching Lin, Wen-Jen Tsay

The reversed order monitoring cusum (ROM) type test with the factor structure is considered. Despite detecting the latest break, we allow the unobserved factors as predictors. We then show that after an appropriate transformation of recursively estimated factors, these estimated factors would not change the asymptotic property of the Brownian motion under regular conditions with suitable ratio of N to T. Monte Carlo simulations show that by using these transformed factors, there is almost no size distortion and the power is promising when we extend the evaluating time period with structure change. We also use the simulation to compare the proposed procedure with the fixed windows approach, and the results reveal that the ROM approach, in general, dominates the fixed window method. We then apply the ROM method to predict monthly growth rate of the U.S. industrial production (IP) and real personal income less transfers (RPI). The results support that the use of the ROM method can improve the out-of-sample forecast as compared to the usual fixed window approach.

C1561: **Point and density exchange rate forecasts using yield curve factors in a time-varying framework**
Presenter: **Anastasia Allayioti**, University of Warwick, United Kingdom

Many papers have offered several explanations for the lack of correlation between exchange rate movements and macroeconomic fundamentals. Recent research attributes this predictive failure to the use of inappropriate proxies for market expectations of future fundamentals. A separate literature has stressed how the presence of structural instabilities might hinder the forecasting ability of many exchange rate models. Taking into account evidence suggesting that the term structure of interest rates reflects expectations of market participants about future economic activity,

we compare the predictive ability of specifications augmented with yield factors relative to the performance of models that use traditional macro-fundamentals and financial predictors. Given that the examined sample period includes a period that was characterized by negative interest rates and considerably narrower yield spreads, the adopted framework explicitly accounts for the existence of an effective lower bound on nominal interest rates. Following a comprehensive in-sample evaluation, we examine the relevance of the competing models in generating accurate point and density forecasts within a set-up that explicitly allows for time-evolving dynamics in both the slope and volatility parameters. The out-of-sample evaluation involves numerous statistical criteria and an assessment of the ability of the models under examination to generate economic value.

C1727: **Constructing dynamic life tables with a single factor model**

Presenter: **David Atance**, University of Alcala, Spain

Co-authors: Eliseo Navarro , Alejandro Balbas

A model is developed to construct dynamic live tables based on the idea that the behavior of whole life table can be explained by a reduced number of factors. In this case these factors are identified with some mortality rates at specific ages. These key mortality rates and model parameters estimates are obtained applying a maximum likelihood criteria under the hypothesis of a binomial distribution of the number of deaths. We develop the single factor version of the model which is implemented to the male and female populations of France and Spain. The model is compared with a set of alternative well-known life tables models. To test the forecasting ability of the model we apply a battery of tests using out of sample data. Despite its simplicity, the outcomes indicate that this model it is not outperformed by other more complex mortality models. Another important advantage of this model is that can it be easily implemented to address some longevity risk linked problems in the context of Solvency II.

Sunday 15.12.2019 08:40 - 10:20 Parallel Session G – CFE-CMStatistics

EO546 Room CLO B01 ANALYSIS OF FUNCTIONAL AND OTHER OBJECT DATA Chair: Alexander Petersen

E0514: **Spatial kriging for replicated anisotropic point processes**
Presenter: **Daniel Gervini**, University of Wisconsin-Milwaukee, United States

In many applications, a temporal point process is observed at different spatial locations, and it is of interest to predict the temporal process at a new location. For example, in the Divvy bicycle-sharing system of the city of Chicago, bike checkout times can be seen as a temporal point process observed every day at different spatial locations, the bike stations, and it is of interest to predict bike demand at potential locations outside the grid, in order to determine where to best place new stations. A spatial kriging approach to this problem will be presented. Unlike common kriging methods that assume isotropy of the spatial field and would not be applicable to the Divvy data, which is clearly anisotropic, the proposed method uses the daily replications of the process at each site to nonparametrically estimate the spatial means and covariances and does not need the isotropy assumption.

E0852: **Wasserstein covariance for vectors of random densities**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Alexander Petersen

Samples of data that consist of probability densities or distributions are encountered in various applications. Once a metric for densities is specified, the Frechet mean or barycenter is typically used to determine the average density. The Wasserstein metric is popular due to its good practical performance and interpretation as an optimal transport metric. Motivated by applications in neuroimaging, we consider data that consist of a p-vector of univariate random densities for each sampling unit. We introduce Wasserstein covariance to quantify the dependency of the component densities and provide corresponding estimators for fixed and diverging p, where the latter corresponds to continuously-indexed densities. Consistency and asymptotic normality are established, while accounting for errors introduced in the unavoidable preparatory density estimation step. The utility of the Wasserstein covariance matrix is demonstrated in applications that include functional connectivity in the brain and the secular evolution of mortality.

E1066: **Statistical challenges for analysis of data in some non-standard spaces**
Presenter: **Stephan Huckemann**, University of Goettingen, Germany

Various kinds of non-Euclidean data, spaces modeling such data, data descriptors, inferential methods and three statistical challenges in this context are introduced. 1) In spaces of globally nonpositive curvature, e.g. in Billera-Holmes-Vogtman (BHV) phylogenetic tree spaces, means of probability distributions are unique but there is the price of "stickiness" to pay: there may be no asymptotic distribution. In contrast, on compact manifolds (already on spheres) uniqueness may be lost, or asymptotic distributions may have arbitrary slow "smeary" rates. 2) In more general spaces, even local uniqueness of geodesics may be lost, such as tropical phylogenetic tree spaces, although they may feature less "stickiness" than BHV spaces. 3) In shape analysis, configurations are studied modulo group actions. Once the configurations are no longer in a Euclidean space (they are Euclidean for classical Kendall shapes), it is unclear how to model directions of internal correlations. Already on a two-torus, the only directions producing nonwinding geodesics are horizontal or vertical. We survey workarounds for higher dimensional torus data, as occur in biomolecule modeling or more general polysphere data, as arise in medical imaging.

E1251: **Non-identifiability of word embeddings, and connections to shape analysis**
Presenter: **Simon Preston**, University of Nottingham, United Kingdom
Co-authors: Karthik Bharath, Rachel Carrington

In statistical shape analysis the shape of a centred configuration of landmarks is the information invariant to orthogonal and scale transformations. We will make the connection between shape analysis and word embeddings. A word embedding is a construction of landmarks such that each landmark represents a distinct word and the relative positions characterise word meaning. Many different models have been proposed for constructing word embeddings, and it is typical to compare different embeddings in terms of how well they perform in word similarity and association tasks. Performance is quantified via a function (of the embedding and some test data) that happens to be invariant to orthogonal and scale transformations of the embedding, and hence can be interpreted as a measure of shape. The criteria optimised when constructing word embeddings are, however, invariant to a wider class of transformations. We will explain this and discuss the problematic consequences for interpreting reported performance of word embeddings.

EO538 Room MAL B02 METHODOLOGICAL DEVELOPMENTS IN MEDICAL STATISTICS AND ITS APPLICATIONS Chair: Chao Huang

E0687: **Comparative intervention scoring for assessing heterogeneity of long-term health system intervention effects**
Presenter: **Jared Huling**, The Ohio State University, United States
Co-authors: Menggang Yu, Maureen Smith

With the growing cost of health care in the United States, the need to improve efficiency and efficacy of the delivery of care has become increasingly urgent. To this end, there have been widespread efforts to design and implement interventions which coordinate the typically fragmented care of complex patients, yet the effectiveness of such interventions in practice has been mixed. A common thread among successful care coordination interventions is the targeting of patients likely to benefit for enrollment, however, there is little guidance toward effectively doing so. We seek to fill this gap by introducing a procedure to estimate personalized scores which characterize differential benefit of long-term health system interventions. As patients tend to respond differently over time, our approach allows the differential effects of an intervention to vary with time and encourage these effects to be more similar for closer time points. We utilize our approach to construct personalized enrollment decision rules for a complex case management intervention in a large health system and demonstrate that the enrollment decision rules result in improvement in health outcomes and care costs.

E0907: **Jointly model longitudinal and semicompeting risks data to accommodate informative dropout and death**
Presenter: **Qiuju Li**, University College London, United Kingdom

In longitudinal studies, both dropout and death can occur during the follow-up and therefore truncate the observation of the longitudinal outcome of interest from a subject. We propose a new likelihood-based approach to accommodating informative dropout and death by jointly modelling the longitudinal outcome and semicompeting event times of dropout and death. Maximum likelihood and Bayesian approaches are used for estimation. Also, since extrapolation beyond death is often not appropriate, it is desirable to obtain the longitudinal outcome profile of a population given being alive. Under the proposed joint modelling framework, the conditional longitudinal profile of being alive can be obtained in a closed form. The proposed methods are illustrated in the application to the HIV Epidemiology Research Study (HERS).

E0941: **Multi-state models for multi-type recurrent events and terminal events with feedback in longitudinal covariates**
Presenter: **Chuoxin Ma**, University of Cambridge, United Kingdom
Co-authors: Jianxin Pan

In cardiovascular disease study, one of the main interest is to investigate the association between risk factors and a series of multi-type recurrent events and terminal events, such as myocardial infarction, stroke and cardiovascular death. When the trajectories of some biomarkers contain

past event feedbacks, existing approaches handling time-dependent covariates in event history analysis can be problematic. We propose a class of multi-state models for the analysis of multi-type recurrent events and terminal events when biomarkers contain past event feedback and are intermittently observed and subject to measurement errors. The competing risk structure and the progressive nature of the multiple events can be well captured by state-specific intensity functions. Both time-varying and constant coefficients can be accommodated. Estimation procedure based on polynomial splines approximation and an extension to the corrected score approach is developed. The consistency and asymptotic normality of the proposed estimators are provided. Simulations show that the naive estimators which either ignore the past event feedback or the measurement errors are biased. Our method achieves better coverage probability of the time-varying/constant coefficients, compared to the naive methods. An application to the data set from the Atherosclerosis Risk in Communities Study is presented.

E0301: **Bayesian analysis of chromosomal interactions in hi-c data using the hidden Markov random field model**
Presenter: **Itunu Osuntoki**, University of Essex, United Kingdom
Co-authors: Andrew Harrison, Hongsheng Dai, Yanchun Bao, Nicolae Radu Zabet

There are different biological methods that have been developed over the years for analysis of the 3D structure of the DNA. Few computational and statistical methods have, however, been developed to analysis data generated using the Hi-C method. We follow statistical methodology to explore the Hi-C data. The Hi-C data is well suited to be analysed using a finite mixture model. The Potts model, a hidden Markov random field model, was employed to analyze the hidden (latent) components. The hidden components are categorised into three; noise, false signal and true signal. Using the Metropolis-within-Gibbs approach to analyze the data, the proposed method was able to detect interactions (short and long range) and false interactions. A large part of the significant interactions that we detect are found within Topological Associated Domains, which is one of the 3D structures known to occur in Hi-C data.

EO448 Room MAL B04 RECENT ADVANCES IN BLIND SOURCE SEPARATION Chair: Sara Taskinen

E0635: **Statistical properties of a blind source separation estimator for complex-valued weakly stationary stochastic processes**
Presenter: **Niko Lietzen**, Aalto University School of Science, Finland
Co-authors: Lauri Viitasaari, Pauliina Ilmonen

Novel asymptotic theory is presented for a blind source separation procedure, in the context of complex-valued signals. In particular, we provide a comprehensive mathematical foundation, applicable for a class of complex-valued blind source separation procedures, for scenarios when the blind source separation estimators have rates of convergence that differ from root-n and when the corresponding estimators have limiting distributions that are not Gaussian. We further investigate the asymptotic behavior of the algorithm for multiple unknown signals extraction (AMUSE) procedure. Under general weakly stationary stochastic processes, obtaining central limit theorem type results is often challenging due to complicated dependency structures. To our knowledge, normalizations that differ from root-n and limiting distributions that differ from the Gaussian distribution have not been previously considered in the blind source separation literature, neither in the real-valued nor the complex-valued case.

E0545: **Blind source separation of graph signals**
Presenter: **Jari Miettinen**, Aalto University, Finland
Co-authors: Sergiy Vorobyov, Esa Ollila

The blind source separation (BSS) problem can be solved either by using non-Gaussianity of the latent components or dependence between the samples. The dependence can be structural such as in time series, spatial data or tensor data, but the focus is on BSS of graph signals which may have more complicated dependencies characterized by graphs and their adjacency matrices. So far, only one BSS method, called GraDe (graph decorrelation), has been designed for this setup. It uses joint approximate diagonalization of graph autocovariance matrices, which are generalizations of autocovariance matrices for time series, and thus GraDe can be seen as generalization of SOBI (second-order blind identification) method for BSS of time series. Modifications of GraDe are suggested, and combining it to non-Gaussianity based methods, FastICA and JADE (joint approximate diagonalization of eigenmatrices), is proposed. In a simulation study, the proposed methods are shown to always achieve at least the performance of the better component (GraDe or FastICA/JADE), and in the case of non-Gaussian graph signals to outperform them both.

E0220: **Spatial blind source separation**
Presenter: **Francois Bachoc**, Universite Paul Sabatier, France
Co-authors: Marc Genton, Klaus Nordhausen, Anne Ruiz-Gazen, Joni Virta

A blind source separation model has been recently suggested for spatial data together with an estimator based on the simultaneous diagonalization of two scatter matrices. The asymptotic properties of this estimator are derived, and a new estimator, based on the joint diagonalization of more than two scatter matrices, is proposed. The limiting properties and merits of the novel estimator are verified in simulation studies. A real data example illustrates the method.

E1088: **Functional independent component analysis**
Presenter: **Germain Van Bever**, Universita de Namur, Belgium
Co-authors: Radka Sabolova, Frank Critchley, Bing Li, Hannu Oja

With the increase in measurement precision, functional data is becoming common practice. Relatively few techniques for analysing such data have been developed, however, and a first step often consists in reducing the dimension via Functional PCA, which amounts to diagonalising the covariance operator. Joint diagonalisation of a pair of scatter functionals has proved useful in many different setups, such has Independent Component Analysis (ICA), Invariant Coordinate Selection (ICS), etc. The Fourth Order Blind Identification procedure is extended to the case of data on a separable Hilbert space (with classical FDA setting being the go-to example). In the finite-dimensional setup, this procedure provides a matrix Gamma such that Gamma X has independent components, if one assumes that the random vector X satisfies $X = \Psi Z$, where Z has independent marginals and Ψ is an invertible mixing matrix. When dealing with distributions on Hilbert spaces, two major problems arise: (i) the notion of marginals is not naturally defined and (ii) the covariance operator is, in general, non invertible. These limitations are tackled by reformulating the problem in a coordinate-free manner and by imposing natural restrictions on the mixing model. The proposed procedure is shown to be Fisher consistent and affine invariant. A sample estimator is provided and its convergence rates are derived. The procedure is amply illustrated on simulated and real datasets.

EO590 Room MAL B18 STATISTICAL METHODS FOR MISSING DATA IN EHR-BASED RESEARCH Chair: Sebastien Haneuse

E0655: **Bayesian nonparametrics for comparative effectiveness research in EHRs**
Presenter: **Michael Daniels**, University of Florida, United States

A Bayesian nonparametric approach is proposed to address both confounding and selection bias for inference using electronic health records (EHRs). Data provenance, the collection of decisions that give rise to the observed data, is modularized and modeled to properly adjust for the selection bias due to missing data. The approach is motivated by a study to assess the long terms effects of bariatric surgery on weight gain.

E1220: **Robust variance estimation when combining MI and IPW for missing data in EHR-based studies**
Presenter: **Tanayott Thaweethai**, Harvard University, United States
Co-authors: Sebastien Haneuse

Due to the complex process by which electronic health records (EHR) are generated and collected, missing data is a huge challenge when conducting

large observational studies using EHR data. Most standard methods to adjust for selection bias due to missing data fail to address the heterogeneous structure of EHR data. We consider a framework that modularizes the data provenance, or the sequence of specific decisions made by patients, health care providers, and the health system in which they interact, that leads to observing complete data in the EHR. Under this framework, one strategy is to combine inverse probability weighting and multiple imputation at different stages to address missingness. We propose an estimator based on this strategy, show that it is consistent and asymptotically Normal, and derive a consistent estimator of the asymptotic variance. Unlike Rubin's standard combining rules for multiple imputation, this variance estimator is robust to model misspecification of the selection, imputation, and analysis models. We demonstrate these methods in an EHR-based study of long-term weight change following bariatric surgery.

E1899: **Electronic health record phenotyping using anchor-positive and unlabeled patients**
Presenter: **Jinbo Chen**, University of Pennsylvania, United States
Co-authors: Lingjiao Zhang, Xiruo Ding, Yanyuan Ma, Naveen Muthu, Imran Ajmal, Jason H Moore, Daniel Herman

Phenotyping patients in electronic health records (EHRs) conventionally relied on algorithms learned from labeled cases and controls. Assigning labels requires manual medical chart review and therefore is an intensive labor. We developed a phenotyping method when identification of gold-standard controls is prohibitive, so a validation set is not available. The method relies on a random subset of cases, which can be specified using an expert-derived anchor variable that has an excellent positive predictive value and sensitivity independent of predictors. Adopting a maximum likelihood approach to efficiently leveraging data from the anchor-labeled cases and unlabeled patients to develop logistic regression phenotyping models, we propose novel statistical methods for internally assessing model calibration and predictive performance measures. Upon identification of an anchor variable by clinical experts that is scalable and transferable to different practices, the approach should facilitate development of scalable, transferable, and practice-specific phenotyping models. Through phenotyping two cardiovascular conditions in Penn Medicine EHRs, we demonstrate that the proposed method enables accurate semi-automated EHR phenotyping with minimal manual labeling and therefore is expected to greatly facilitate EHR clinical decision support and research.

E1957: **Handling missing data in propensity score analyses of electronic health record data**
Presenter: **Elizabeth Williamson**, London School of Hygiene and Tropical Medicine, United Kingdom

Missing data are a ubiquitous problem in medical research. When the data being used to address the research question are not collected for the purpose of research, the extent of missing data tends to be greater. Further, the mechanisms by which data become missing in EHR are likely to be different from those operating in more traditional study designs. The focus will be on studies where the main aim is to address a question of causal inference, for example estimating the comparative effect of two drugs on a health outcome. Particularly in settings using routinely collected health data, such as electronic health record (EHR) data, propensity score analysis is frequently applied; this analysis framework raises additional issues related to missing data. The way in which missing data is handled in these studies can lead to imprecise estimates and/or bias. Various commonly used missing data methods will be discussed, and the plausibility of the assumptions underlying these methods will be explored in studies using data from electronic health records. It will demonstrate that some missing data methods commonly described as ad-hoc can produce valid statistical inference under specific assumptions which can sometimes be plausible in these settings.

EO761 Room MAL B20 CAUSAL INFERENCE IN FACTORIAL EXPERIMENTS Chair: Peng Ding

E0724: **Randomization tests for weak null hypotheses**
Presenter: **Peng Ding**, University of California, Berkeley, United States

The Fisher randomization test (FRT) is applicable for any test statistic, under a sharp null hypothesis that can recover all missing potential outcomes. However, it is often of interest to test a weak null hypothesis that the treatment does not affect the units on average. To use the FRT for a weak null hypothesis, we must address two issues. First, we need to impute the missing potential outcomes although the weak null hypothesis cannot determine all of them. Second, we need to choose an appropriate test statistic. For a general weak null hypothesis, we propose an approach to imputing missing potential outcomes under a compatible sharp null hypothesis. With this imputation scheme, we advocate using a studentized statistic. The resulting FRT has multiple desirable features. First, it is model-free. Second, it is finite-sample exact under the sharp null hypothesis that we use to impute the potential outcomes. Third, it preserves correct large-sample type I errors under the weak null hypothesis of interest. Therefore, our FRT is agnostic to treatment effect heterogeneity. We establish a unified theory for general factorial experiments. We also extend it to stratified and clustered experiments.

E0727: **Rerandomization in 2^K factorial experiments**
Presenter: **Xinran Li**, University of Illinois Urbana-Champaign, United States

With many pretreatment covariates and treatment factors, the classical factorial experiment often fails to balance covariates across multiple factorial effects simultaneously. Therefore, it is intuitive to restrict the randomization of the treatment factors to satisfy certain covariate balance criteria, possibly conforming to the tiers of factorial effects and covariates based on their relative importances. This is rerandomization in factorial experiments. We study the asymptotic properties of this experimental design under the randomization inference framework without imposing any distributional or modeling assumptions of the covariates and outcomes. We derive the joint asymptotic sampling distribution of the usual estimators of the factorial effects, and show that it is symmetric, unimodal and more concentrated at the true factorial effects under rerandomization than under the classical factorial experiment. We quantify this advantage of rerandomization using the notions of central convex unimodality and peakedness of the joint asymptotic sampling distribution. We also construct conservative large-sample confidence sets for the factorial effects.

E1407: **Causal interaction in factorial experiments: Application to conjoint analysis**
Presenter: **Naoki Egami**, Princeton University, United States
Co-authors: Kosuke Imai

Causal interaction is studied in factorial experiments, in which several factors, each with multiple levels, are randomized to form a large number of possible treatment combinations. Examples of such experiments include conjoint analysis, which is often used by social scientists to analyze multidimensional preferences in a population. To characterize the structure of causal interaction in factorial experiments, we propose a new causal interaction effect, called the average marginal interaction effect (AMIE). Unlike the conventional interaction effect, the relative magnitude of the AMIE does not depend on the choice of baseline conditions, making its interpretation intuitive even for higher-order interactions. We show that the AMIE can be nonparametrically estimated using ANOVA regression with weighted zero-sum constraints. Because the AMIEs are invariant to the choice of baseline conditions, we directly regularize them by collapsing levels and selecting factors within a penalized ANOVA framework. This regularized estimation procedure reduces false discovery rate and further facilitates interpretation. Finally, we apply the proposed methodology to the conjoint analysis of ethnic voting behavior in Africa and find clear patterns of causal interaction between politicians ethnicity and their prior records.

E1472: **Regression-adjusted average treatment effect estimates in stratified and sequentially randomized experiments**
Presenter: **Hanzhong Liu**, Tsinghua University, China

Stratified and sequentially randomized experiments are widely used. Baseline covariates are often collected for each unit. Linear regressions are sometimes used to adjust minor imbalances of covariates in the treatment and control group. Asymptotic properties of regression adjustment in stratified and sequentially randomized experiments are studied under randomization-based inference. We allow both the number of strata and their sizes to be arbitrary, provided the total number of experimental units tends to infinity and each stratum has at least two treated and two control units. Under slightly stronger, we re-establish the finite population CLT for a stratified random sample. We prove that, under mild conditions, both

the stratified difference-in-means and the regression-adjusted average treatment effect estimator are consistent and asymptotically normal. The asymptotic variance of the latter is no greater, and is typically lesser than that of the former when the proportion of treated units is asymptotically the same across strata or the number of stratum is bounded. The improvement depends on the extent to which the within-strata variation of the potential outcomes can be explained by the covariates. We also provide conservative variance estimators to construct large-sample confidence intervals for the average treatment effect, which are consistent if and only if the stratum-specific treatment effect is constant. Simulations and empirical illustration are provided.

EO540 Room MAL B35 INFERENCE METHODS IN SURVIVAL ANALYSIS Chair: Dennis Dobler

E0302: **Studying the influence of time-varying treatment regimes on a time-to-event outcome using Danish registry data**
Presenter: **Sarah Friedrich**, University Medical Center Goettingen, Germany

In a Danish nationwide study in diabetes patients based on registry data, the aim was to compare the effect of different treatment regimes on a subsequent time-to-event outcome. However, definition of treatment regimes is complicated due to patients switching back and forth between treatments. Moreover, the competing risk of death further complicates analyses. We employ a nested case-control design which allows for the definition of comparable treatment histories for cases and controls. Gaining a causal interpretation of the effect estimates is not straightforward in this setting due to the complicated treatment histories, which require advanced methods of causal inference (like targeted learning) in a time-continuous framework. We discuss the strengths and weaknesses of the approach and give an outlook on the methods and developments needed to gain a causal interpretation in such a setting.

E0320: **Inference for factorial designs in survival models**
Presenter: **Marc Ditzhaus**, Technical University of Dortmund, Germany
Co-authors: Markus Pauly, Arnold Janssen

Inference procedures are proposed for general nonparametric factorial survival designs with possibly right-censored or left-truncated time-to-event data. Similar to additive models, null hypotheses are formulated in terms of cumulative hazards. Thereby, deviations are measured in terms of quadratic forms in Nelson-Aalen-type integrals. In contrast to existing approaches, this allows us to work without restrictive model assumptions on the survival, censoring or truncation times. For a distribution-free application of the method, permutation and wild bootstrapping are suggested. The resulting procedures' asymptotic validity, as well as their consistency, are proven, and their small sample performances are analyzed in extensive simulations. Finally, their applicability is demonstrated by an illustrative data analysis.

E0442: **State occupation probabilities in non-Markov models**
Presenter: **Morten Overgaard**, Aarhus University, Denmark

In multi-state models, state occupation probabilities may well be of interest. When the multi-state process has the Markov property, estimates of the state occupation probabilities can be based on the Aalen–Johansen estimates of the transition probabilities. This estimate is consistent under an independent censoring assumption. The estimate remains consistent when the multi-state process does not have the Markov property. We will take a look at why that is. The approach will be through seeing the transition probabilities as interval functions and studying certain additive and multiplicative transforms related to these interval functions. Under a bounded variation requirement, an appropriate expression of the state occupation probability can then be given in terms of the initial distribution and a product integral of the transition hazards even without the Markov property. This expression matches what is estimated by the Aalen–Johansen-based estimator of the state occupation probabilities under the independent censoring assumption thus establishing the claim.

E1286: **Collapsible Cox-regression and non-collapsible Aalen additive hazards regression**
Presenter: **Sven Ove Samuelsen**, University of Oslo, Norway

It is well-known that the additive hazards model is collapsible, in the sense that when omitting one covariate from a model with two independent covariates, the marginal model is still an additive hazards model with the same regression coefficient. In contrast, for the proportional hazards model under the same covariate assumption, the marginal model is no longer a proportional hazards model and is not collapsible. These results, however, relate to the model specification and not to the regression estimates. We point out that if covariates in risk sets at all event times are independent then both Cox and Aalen regression estimates are collapsible, in the sense that there is no systematic change in the parameter estimates. Vice-versa, if this assumption fails, then the estimates will change systematically both for Cox and Aalen regression. In particular, if the data are generated by an Aalen model with censoring independent of covariates both Cox and Aalen regression is collapsible, but if generated by a proportional hazards model neither estimators are. We will also discuss settings where survival times are generated by proportional hazards models with censoring and truncation patterns providing uncorrelated covariates and hence collapsible Cox and Aalen regression estimates.

EO066 Room MAL B36 RANDOM FORESTS AND APPLICATIONS Chair: Efoevi Angelo Koudou

E1073: **Uncertain trees: Dealing with uncertain inputs in regression trees**
Presenter: **Myriam Tami**, CentraleSupelec, France
Co-authors: Marianne Clausel, Emilie Devijver, Eric Gaussier, Sami Alkhoury

Tree-based ensemble methods, as Random Forests and Gradient Boosted Trees, have been successfully used for regression problems in many applications and research studies. Furthermore, these methods have been extended in order to deal with uncertainty in the output variable, using for example a quantile loss in Random Forests. We study a generalization of regression trees, referred to as uncertain trees, that deals with uncertainties in the input variables. By doing so, one no longer assumes that an observation lies into a single region of the regression tree, but rather that it is associated with each region with a certain probability that depends on the distance between the observation and the region. This generalization, which differs from soft decision trees, raises several questions and we show, in particular, that uncertain trees are consistent. Experiments conducted on several data sets further illustrate the good behavior of uncertain trees.

E1493: **Frechet random forests**
Presenter: **Louis Capitaine**, Bordeaux University INSERM, France
Co-authors: Rodolphe Thiebaut, Robin Genuer

Random forests are a statistical learning method widely used in many areas of scientific research, essentially for its ability to learn complex relationship between input and output variables, and also its capacity to handle high-dimensional data. However, data are increasingly complex with repeated measures of omics, images leading to shapes, curves... The random forests method is not specifically tailored for them. We introduce Frechet trees and Frechet random forests, which allow us to manage data for which input and output variables take values in general metric spaces (which can be unordered). To this end, a new way of splitting the nodes of trees is introduced and the prediction procedures of trees and forests are generalized. Then, random forests out-of-bag error and variable importance score are naturally adapted. Finally, the method is studied in the special case of regression on curve shapes, both within a simulation study and a real dataset from an HIV vaccine trial.

E1533: **Short term wind power forecasting and ramp alert forecasting using random forest**
Presenter: **Dione Mamadou**, CREST (ENSAE ParisTech), France

The breaking of purchase obligation contracts with the energy transition law defined by the French State involves the sale of wind power on the market. For this sale, it will be necessary to announce in advance the amount of electricity to produce. So, we need production forecasts. We

propose a random forest model, and then, an aggregation of the forecasts of several wind farms to reduce the uncertainties and alerts of ramps to prevent the periods of large forecasting errors.

E1545: **Random forests: A survey**
Presenter: **Efoevi Angelo Koudou**, IECL CNRS /Universite de Lorraine, France
A summary of the fundamentals of Random forests is presented, as well as a few instances of application of this method in the recent literature.

EO564 Room MAL G13 LOCAL EMPIRICAL MEASURES AND NONPARAMETRIC STATISTICS Chair: Davit Varron

E1232: **Auxiliary information empirical processes and their bootstrap**
Presenter: **Philippe Berthet**, Toulouse University, France
In nonparametric statistics, when some information on the unknown distribution is available but not enough to justify a parametric or semi parametric model, a crucial question is: how to exploit the auxiliary information? The iterative raking-ratio procedure consists in changing the weights of the empirical measure to fit some true requirement, typically a discrete marginal distribution as in survey analysis, and then iterate to fit the next requirement. Kullback first proved its convergence. We obtain asymptotic, and non asymptotic results - in sample size - for a fixed number of iterations, then infinite. The main feature is the uniform decrease of bias, variance, covariance and quadratic risk over large classes of linear estimators - measured through the empirical process indexed by functions - and the closed form expression of the limiting Gaussian process - that is no more a Brownian Bridge. Uniform Berry-Esseen results and concentration probability bounds follow by coupling the raking-ratio empirical process to its limiting process. In order to evaluate the decrease of risk, it is natural to bootstrap the raked statistics. Theoretical results on the weighted bootstrap will also be presented, when each random weight depends on the sample point. By using a joint strong approximation we provide a size of Monte-Carlo bootstraps allowing a mathematical control of the unavoidable bias, variance and distribution function distortion for regular statistics.

E1871: **A class of conditional empirical measures**
Presenter: **Davit Varron**, University of Franche-Comte, France
The focus is on a class of random point processes that share properties with empirical measures when conditioned to another exogenous random phenomenon. We investigate the validity of Glivenko-Cantelli and Donsker theorems for such random measures. In this setup, we prove that the usual conditions on uniform entropy numbers are strong enough to derive these two theorems. Some applications of these results are also presented in the framework of extreme value theory and nearest-neighbour rules.

E1873: **Excess risk concentration for M-estimators**
Presenter: **Adrien Saumard**, Crest-Ensai, France
General functionals of M-estimators can be represented as maximizers of some local empirical processes indexed by some subsets of functions in the model. Among all functionals that can be considered, the excess risk plays a central role because it is an intrinsic measure of the performance of M-estimators. By using these representation formulas, one can look at a finer scale than the (minimax) rates of convergence, corresponding to concentration rates. We tackle the case of a convex and regular loss, by applying a concavity argument combined with a linearization of the contrast. By specializing the model structure, we also address the remaining problem of a sharp computation of the mean of the excess risk.

EO188 Room MAL G14 PRESENT-DAY DATA ANALYSIS CHALLENGES MEET BAYES Chair: Mario Peruggia

E0323: **Diversity and precision of Bayesian Mallows to learn preferences from clicking data**
Presenter: **Arnoldo Frigessi**, University of Oslo, Norway
Clicking data contain user preference information and can be used to produce personalized recommendations in web-based applications. We propose the Bayesian Mallows for Clicking Data method, which augments clicking data into compatible full ranking vectors. User preferences are learned using a Mallows ranking model. Bayesian inference leads to interpretable uncertainties of each individual recommendation. With a simulation study and a data example, we demonstrate that compared to state-of-the-art matrix factorization, our method makes personalized recommendations with similar accuracy, while achieving higher level of diversity, and producing interpretable and actionable uncertainty estimation. We discuss computationally efficient model approximations.

E0345: **Locally stationary processes and their application to climate modeling**
Presenter: **Peter Craigmile**, The Ohio State University, United States
In the analysis of climate it is common to build non-stationary spatio-temporal processes, often based on assuming a random walk behavior over time for the error process. Random walk models may be a poor description for the temporal dynamics, leading to inaccurate uncertainty quantification. Likewise, assuming stationarity in time may also not be a reasonable assumption, especially under climate change. Based on ongoing research, we present a class of time-varying processes that are stationary in space, but locally stationary in time. We demonstrate how to carefully parameterize the time-varying model parameters in terms of a transformation of basis functions. We present some properties of parameter estimates when the process is observed at a finite collection of spatial locations, and apply our methodology to a Bayesian spatio-temporal climate analysis.

E0428: **Hierarchical Hidden Markov models for response time data**
Presenter: **Deborah Kunkel**, Clemson University, United States
Psychological data, particularly measurements obtained sequentially in experiments designed to test theories of human cognition, are often treated as independent and identically distributed samples from a single distribution that describes the cognitive process. This assumption is made for mathematical and analytic convenience; it is widely appreciated that such data are in fact mixtures from two or more processes, a subset of which are associated with the cognitive process of interest. Our modeling framework describes response times (RTs) as arising from a mixture of three distinct distributions. Transitions across the distributions are governed by a hidden Markov structure whose states produce either fast, average, or slow RTs. This process is nested within a second Hidden Markov structure, producing an 'environment' process that allows the distribution of the response status to evolve due to factors such as fatigue and distractions. We performed a detection experiment designed to elicit responses under three environments that mimic the external conditions thought to influence latent statuses. We present our hierarchical model and demonstrate its fit on the experimental data.

E0986: **Assessing forecasts from data of different dimension**
Presenter: **Catherine Forbes**, Monash University, Australia
Co-authors: Worapree Ole Maneesoonthorn
Consider the following competitive scenario. Modeller One (M1) produces a univariate forecast distribution for a target future observation using a simple model depending only on the historical trajectory of the univariate target series. Modeller Two (M2) produces a multivariate forecast distribution, where one of the variables is the same target variable. Who produces the best forecasts? How are we to assess the forecast distributions produced by the competitors? The aim is first to discuss situations where the competitive scenario reasonably occurs, and why a naive comparisons may be problematic. We offer both theoretical and practical suggestions for the construction of appropriate scoring rules to compare the performance of the resulting forecast distributions. Finally, we use the insights gained to combine the competing forecasts into a single target forecast distribution, in case M1 and M2 decide that they want to collaborate.

EO634 Room MAL G15 RECENT DEVELOPMENTS IN BAYESIAN COMPUTATION Chair: Matti Vihola

E0283: **Finding our way in the dark: Approximate MCMC for approximate Bayesian methods**
Presenter: **Radu Craiu**, University of Toronto, Canada
Co-authors: Evgeny Levi

With larger amounts of data at their disposal, scientists are emboldened to tackle complex questions that require sophisticated statistical models. It is not unusual for the latter to have likelihood functions that elude analytical formulations. Even under such adversity, when one can simulate from the sampling distribution, Bayesian analysis can be conducted using approximate methods such as Approximate Bayesian Computation (ABC) or Bayesian Synthetic Likelihood (BSL). A significant drawback of these methods is that the number of required simulations can be prohibitively large, thus severely limiting their scope. We design perturbed MCMC samplers that can be used within the ABC and BSL paradigms to significantly accelerate computation while maintaining control on computational efficiency. The proposed strategy relies on recycling samples from the chain's past. The algorithmic design is supported by a theoretical analysis while practical performance is examined via a series of simulation examples and data analyses.

E0559: **Ensemble MCMC: Accelerating pseudo-marginal MCMC for state space models using the ensemble Kalman filter**
Presenter: **Richard Everitt**, University of Warwick, United Kingdom

Particle Markov chain Monte Carlo (pMCMC) is now a popular method for performing Bayesian statistical inference on challenging state space models (SSMs) with unknown static parameters. It uses a particle filter (PF) at each iteration of an MCMC algorithm to unbiasedly estimate the likelihood for a given static parameter value. However, pMCMC can be computationally intensive when a large number of particles in the PF is required, such as when the data is highly informative, the model is misspecified and/or the time series is long. In this paper we exploit the ensemble Kalman filter (EnKF) developed in the data assimilation literature to speed up pMCMC. We replace the unbiased PF likelihood with the biased EnKF likelihood estimate within MCMC to sample over the space of the static parameter. On a wide class of different non-linear SSM models, we demonstrate that our new ensemble MCMC (eMCMC) method can significantly reduce the computational cost whilst maintaining reasonable accuracy. We also propose several extensions of the vanilla eMCMC algorithm to further improve computational efficiency and allow for approximate posterior hidden state inference.

E1077: **Parallelising particle filters with butterfly interactions**
Presenter: **Kari Heine**, University of Bath, United Kingdom
Co-authors: Nick Whiteley, Ali Taylan Cemgil

Bootstrap particle filter (BPF) is the cornerstone of many algorithms used for solving generally intractable inference problems with Hidden Markov models. The long term stability of BPF arises from particle interactions that typically make parallel implementations of BPF nontrivial. We propose a method whereby the particle interaction is done in several stages. With the proposed method, full interaction can be accomplished even if we allow only pairwise communications between processing elements at each stage. We show that our method preserves the consistency and the long term stability of the BPF, although our analysis suggest that the constraints on the stagewise interactions introduce error leading to a lower convergence rate than standard Monte Carlo. The proposed method also suggests a new, more flexible, adaptive resampling scheme, which according to our numerical experiments is the method of choice, displaying a notable gain in efficiency in certain parallel computing scenarios.

E1141: **Geometric MCMC for infinite-dimensional inverse problems**
Presenter: **Alexandros Beskos**, University College London, United Kingdom
Co-authors: Andrew Stuart, Mark Girolami, Patrick Farell, Shiwei Lan

Bayesian Inverse Problems often involve sampling posteriors on infinite-dimensional spaces. Traditional MCMC algorithms are characterized by deteriorating mixing times upon mesh-refinement, when the finite-dimensional approximations become more accurate. Such methods are forced to reduce step-sizes as the discretization gets finer, thus are expensive as a function of dimension. Recently, a new class of MCMC methods with mesh-independent convergence times has emerged. However, few of them take into account the geometry of the posterior. At the same time, geometric MCMC algorithms have been found to be powerful in exploring complicated distributions that deviate from elliptic Gaussian laws, but are computationally intractable for models defined in infinite-dimensions. We combine geometric methods on finite-dimensional subspaces with mesh-independent infinite-dimensional approaches. The objective is to speed up MCMC mixing, without significantly increasing the computational cost per-step. This is achieved by using ideas from geometric MCMC to probe the complex structure of an intrinsic finite-dimensional subspace where most data information concentrates, while retaining robust mixing times as the dimension grows by using pCN-like methods in the complementary subspace. The resulting algorithms are demonstrated in the context of 3 challenging Inverse Problems and can exhibit up to two orders of magnitude improvement in sampling efficiency when compared with pCN.

EO286 Room MAL G16 DEPENDENCE MEASURES Chair: Sebastian Fuchs

E0507: **The local Gaussian correlation with applications to finance**
Presenter: **Baard Stoeve**, University of Bergen, Norway

The local Gaussian correlation (LGC) is discussed. LGC is a dependence measure capable of describing non-linear relationships. We will use this measure to study asymmetric dependence structures between financial returns, in particular give evidence of increased dependence during bear markets. We will also present the use of the LGC in the portfolio optimization problem.

E0223: **Compatibility and attainability of matrices of correlation-based measures of concordance**
Presenter: **Takaaki Koike**, University of Waterloo, Canada
Co-authors: Marius Hofert

Measures of concordance have been widely used to summarize non-linear dependence among random variables, which Pearson's correlation coefficient cannot capture. However, popular measures of concordance, such as Spearman's rho and Blomqvist's beta, appear as classical correlations of transformed random variables. We characterize a whole class of such concordance measures arising from correlations of transformed random variables, which includes Spearman's rho, Blomqvist's beta and van der Waerden's coefficient as special cases. Compatibility and attainability of square matrices with entries given by such measures are studied, that is, whether a given square matrix can be realized as a matrix of such pairwise measures of concordance of some random vector, and how such a random vector can be constructed.

E0595: **A unified approach for tail-dependence and concordance measures, with some new indices**
Presenter: **Claudio Giovanni Borroni**, University of Milano - Bicocca, Italy
Co-authors: Lucio De Capitani

Some applications need indices measuring the limiting amount of dependence in the upper and lower corners of a copula. Classical Sibuya's tail-dependence coefficient is based on the limit of the diagonal section of the copula, something which is often regarded as too restrictive, because the index does not account for dependence along routes to the corners other than the main diagonal. Essentially, such a limitation can be ascribed to the link of Sibuya's coefficient to a rather elementary dependence measure, Blomqvist's beta. In the literature some attempts are then made to build other indices upon more sophisticated dependence measures, even if the choice is often more motivated by the mathematical tractability of the measure, such as for Spearman's rho, rather than by its intrinsic properties. We aim, first, at building a systematic approach to link tail-dependence indices to association measures or, more specifically, to measures of concordance. Secondly, we evaluate the use of a class of highly informative

measures, which can clearly distinguish among different cases of a lack of association, independence or reflection invariance. The elements of that class, sometimes referred as mutual, include Kendall's tau other non-trivial measures. Thirdly, estimation issues are addressed.

E0597: **On a class of measures of concordance for bivariate copulas**
Presenter: **Sebastian Fuchs**, TU Dortmund, Germany

The purpose is to study a wide class of measures of concordance for bivariate copulas in which each element $kappa_A$ is generated by a fixed but arbitrary copula A. This class contains Spearman's rho, which is induced by the independence copula, and Gini's gamma, which is induced by the comonotonicity copula. Our approach sheds some new light on Spearman's rho and Gini's gamma and allows for the construction of other meaningful measures of concordance focussing on different facets of dependence between two random variables. In particular, we introduce a measure of concordance that is induced by the countermonotonicity copula. For all measures of concordance in this class, we propose a general construction of a sample version which is based on the empirical copula, and we show that these estimators are asymptotically normally distributed. For Spearman's rho and Gini's gamma it turns out that our sample versions coincide with the usual ones. Moreover, the pointwise order on copulas induces an order relation on this class of measures of concordance. It turns out that, for every copulas C which is left tail decreasing and right tail increasing, the values $\kappa_A(C)$ are decreasing when A is increasing and are bounded below by the corresponding values of Kendall's tau.

EO148 Room CLO 101 LEARNING FOR HIGH-DIMENSIONAL DATA WITH COMPLEX DEPENDENCE Chair: Tetyana Pavlenko

E0582: **Feature selection for sparse mixtures with dependence structure**
Presenter: **Annika Tillander**, Linkoping University, Sweden
Co-authors: Tetyana Pavlenko

Including irrelevant features may deteriorate the classification accuracy and for high-dimensional data, such as e.g. gene expressions, few of the features are expected to be relevant for any given classification problem, hence the need to identify informative features. This is a challenging task when informative features are rare and weak. Accounting for the relation between features can improve the chance to identify the relevant information and this leads to block-wise feature selection. A three-step method is suggested where the first step is to learn the structure between features, the second step is to estimate a measure of information strength, and the third step is a thresholding procedure. For single feature selection, the Higher Criticism is a well-known thresholding method that is optimally adaptive i.e. performs well without knowledge of the sparsity and weakness parameters. This method is extended to handle thresholding for blocks of features. Further, it is compared to other goodness-of-fit tests based on sup-functionals of weighted empirical process for thresholding. The relevance and benefits of feature selection for classification problems is demonstrated using both simulation and real data.

E0810: **Asymptotic robustness for error rate of 2 group discriminant analysis for large dimensional case**
Presenter: **Takayuki Yamada**, Shimane university, Japan
Co-authors: Tetsuro Sakurai, Yasunori Fujikoshi

The focus is on the problems for 2-groups linear discriminant analysis for high-dimensional data when the covariance matrices are equal. Firstly, we show that the asymptotic approximation of the error rate under non-normality as the dimension and sample size go to infinity together. An asymptotic estimator for the error rate is also obtained under the above asymptotic framework. A small-scaled simulation is carried out to confirm the precision of the approximation. We also show an asymptotic approximation for the error rate of the unified type discriminant statistic which includes linear discriminant function and quadratic discriminant function.

E1298: **A robustness evaluation of some Bayesian testing problems**
Presenter: **Lukas Arnroth**, Uppsala University, Sweden
Co-authors: M Rauf Ahmad

The aim is to evaluate certain Bayesian tests for robustness under general classes of distributions. Whereas there is an ample amount of literature on the study of robustness for tests constructed under Neyman-Pearson theory, there is a serious lack of such study for Bayesian testing theory. To prepare the grounds for a solid case, we begin with Bayesian univariate mean tests, which is parallel to the frequentist t-tests. We evaluate them for robustness under a wider class of distributions, keeping a reasonable basis for the priors. Preliminary results are promising. We then extend it to the multivariate testing problems using potential alternatives to multivariate normal distribution, such as the elliptical class. Comparisons with corresponding frequentist tests, on grounds as similar as possible, are discussed. Finally, we extend the study in other directions, e.g., linear models.

E1293: **Optimal detection of sparse mixtures with applications to high-dimensional classification**
Presenter: **Tetyana Pavlenko**, KTH Royal Institute of Technology, Sweden

The focus is on the sparse mixture detection problem for a general non-Gaussian case. We present a class of tests procedures and provide an explicit characterization of the optimal detection boundary under mild regularity conditions. Applications of the obtained results are demonstrated for the adaptive feature selection in high-dimensional classification.

EO340 Room CLO 102 NOVEL TIME SERIES MODELS AND APPLICATIONS Chair: Hernando Ombao

E1713: **Flexible and robust mixed Poisson INGARCH models**
Presenter: **Wagner Barreto-Souza**, Universidade Federal de Minas Gerais, Brazil
Co-authors: Rodrigo Silva

A general class of INteger-valued Generalized AutoRegressive Conditional Heteroskedastic (INGARCH) models is proposed which is based on a flexible family of mixed Poisson (MP) distributions. The proposed class of count time series models contains the negative binomial (NB) INGARCH process as a particular case, and open the possibility to introduce new models such as the Poisson-inverse Gaussian (PIG) and Poisson generalized hyperbolic secant processes. In particular, the PIGINGARCH model is an interesting and robust alternative to the NB model. We explore first-order and second-order stationary properties of our MPINGARCH models and provide expressions for the autocorrelation function and mean and variance marginals. Conditions to ensure strict stationarity and ergodicity properties for our class of INGARCH models are established. We propose an Expectation-Maximization algorithm to estimate the parameters and obtain the associated information matrix. Further, we discuss two additional estimation methods. Monte Carlo simulation studies are considered to evaluate the finite-sample performance of the proposed estimators. We illustrate the flexibility and robustness of the MPINGARCH models through two real-data applications about number of cases of Escherichia coli and Campylobacter infections.

E1927: **Dynamic functional connectivity: A sparse group fused lasso approach**
Presenter: **David Degras**, University of Massachusetts Boston, United States

A novel approach is presented to assess dynamic functional connectivity (DFC) in neuroimaging data. Modeling brain signals as piecewise structural vector autoregressive (SVAR) processes, the method decomposes the time range of the data into intervals of homogeneous functional connectivity (FC). The piecewise SVAR model can be rapidly fitted to data (typically in minutes) thanks to an efficient implementation of sparse group fused lasso (SGFL). To handle the high dimension of the parameter space, SGFL makes two sparsity assumptions: (i) at each time point, there are only few nonzero regression coefficients (i.e. the number of functional connections between pairs of brain regions is small), and (ii) regression coefficients only change infrequently over time, i.e. regime changes in FC are relatively rare. Enforcing these assumptions in model fitting produces a challenging problem of nonsmooth convex optimization, which we solve with a novel hybrid algorithm that combines block coordinate descent,

forward-backward algorithms, iterative soft thresholding, and subgradient methods at different levels of optimization. A numerical comparison of the hybrid approach with state-of-the-art optimization procedures is presented, as well as applications of SGFL to resting-state fMRI and EEG data.

E1941: **Fast and efficient parameter estimation in time series and random fields**
Presenter: **Adam Sykulski**, Lancaster University, United Kingdom
Co-authors: Arthur Guillaumin, Sofia Olhede
Balancing computational and statistical efficiency is a modern challenge of statistical inference. We discuss new methods for parameter estimation in time series and random fields which addresses this very challenge. Specifically, we propose a class of new pseudo-likelihood estimators which are order $N \log N$ to compute, and yield parameter estimates with optimal $\sqrt{N}$ convergence under weaker assumptions than alternative methods. The procedure is inspired from the Whittle likelihood, and as thus is based in the frequency domain, but we make important bias corrections to vastly improve performance. We also extend the procedure to include missing data and irregular spatial shapes, as well as non-linear, non-stationary and anisotropic stochastic processes. We demonstrate the applicability of our techniques to massive datasets across oceanography and the geosciences.

E1948: **Change points methods for high dimensional time series networks**
Presenter: **Ivor Cribben**, Alberta School of Business, Canada
Original statistical methodology on the evolving interdependencies between high-dimensional multivariate time series is developed. Specifically, we introduce a data-driven method which detects change points in the network summary statistics of a (very high dimensional) multivariate time series, with each component of the time series represented by a node in the network. The novel method allows for estimation of both the time of change in the network summary statistics without prior knowledge of the number or location of the change points. We also propose a new multiple change point segmentation method. We show the improvement of our method over classical binary segmentation methods. We apply these methods to various simulated high dimensional data sets as well as to a resting state functional magnetic resonance imaging (fMRI) data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The method allows us to characterize the large scale resting state dynamic brain networks that are related to Alzheimer's disease.

EO180 Room Court SEMIPARAMETRIC AND MIXTURE MODELS AND THEIR USE FOR FRACTIONAL IMPUTATION Chair: Changbao Wu

E0416: **Maximum pairwise-rank-likelihood-based inference for the semiparametric transformation model**
Presenter: **Pengfei Li**, University of Waterloo, Canada
The linear transformation model is studied in the most general setup. This model includes many important and popular models in statistics and econometrics as special cases. Although it has been studied for many years, the methods in the literature either are based on kernel-smoothing techniques or make use of only the ranks of the responses in the estimation of the parametric components. The former approach needs a tuning parameter, which is not easily optimally specified in practice; and the latter approach may be less accurate and computationally expensive. We propose two methods: a pairwise rank likelihood method and a score-function-based method based on this pairwise rank likelihood. Our methods estimate all the unknown parameters in the linear transformation model, and we explore the theoretical properties of our proposed estimators. Via extensive numerical studies, we demonstrate that our methods are appealing in that the estimators are not only robust to the distribution of the random errors but also in many cases more accurate than those of the existing methods.

E0440: **Learning finite mixture models by minimum Wasserstein distance estimator**
Presenter: **Jiahua Chen**, University of British Columbia, Canada
When a population exhibits a level of heterogeneity, finite mixture models provide an easy interpretation: the population is made of several homogeneous subpopulations all from a parametric distribution family. As early as in 1894, Pearson used a two-component Gaussian mixture to fit a crab data set, suggesting the existence of two subspecies. Pearson used the method of moments likely for the ease of numerical computation. Contemporary practice in statistics favours the learning by maximum likelihood for statistical efficiency and the convenient EM-algorithm. The maximum likelihood estimator (MLE) searches for a distribution in the assumed distribution family that attains the minimum Kullback-Leibler divergence from the empirical distribution. Such minimum distance principle can be applied to learn mixtures based on any distances between two distributions. In the machine learning community, the Wasserstein distance has drawn increased attention for its intuitive geometric interpretations and it is successfully employed in many new applications. We study the minimum Wasserstein distance estimator for learning finite Gaussian mixtures. We establish its statistical consistency and demonstrate its superior performances in some applications compared with a penalized version of MLE as the MLE is not well defined for finite Gaussian mixtures.

E1371: **General purpose multiply robust data integration procedure for combining probability and non-probability samples**
Presenter: **David Haziza**, Universita de Montraal, Canada
Co-authors: Sixia Chen
Recently, there has been an increased interest in combining probability and non-probability samples. Non-probability samples are cheaper and quicker to conduct. However, the resulting estimators are vulnerable to bias as the selection probabilities are unknown. To adjust for the potential bias, estimation procedures based on parametric or nonparametric models have been discussed in the literature. The validity of the resulting estimators depends on the validity of the underlying model. Nonparametric approaches may suffer from the curse of dimensionality and loss of efficiency. We propose a data integration approach by combining multiple outcome regression models and propensity score models. The proposed approach can be used for estimating general parameters including totals, means, distribution functions and percentiles. The resulting estimators are multiply-robust in the sense that they remain consistent if all but one model are misspecified. The asymptotic properties of point and variance estimators are established. The results from a simulation study shows the benefits of the proposed method in terms of bias and efficiency.

E1368: **Semiparametric fractional imputation using conditional Gaussian mixture models**
Presenter: **Jae Kwang Kim**, Iowa State University, United States
Imputation is a popular technique for handling item nonresponse in survey sampling. Semiparametric imputation is a robust imputation method that is based on a flexible model where the number of parameters in the model can increase with the sample size. Gaussian mixture model (GMM) imputation is one of the examples of the semiparametric imputation. We propose another semiparametric imputation based on a more flexible model assumption than the GMM. In the proposed mixture model, we still assume a Gaussian model for the conditional distribution of the study variable given the auxiliary variable, but the marginal distribution of the auxiliary variable is not necessarily Gaussian. We show that the proposed mixture model based on the conditional Gaussian mixture achieves a lower approximation error bound to any unknown target density than the GMM in terms of the Kullback-Leibler divergence measure. The proposed method is applicable to high dimensional covariate problems by including a penalty function in the conditional log-likelihood function. The parameter estimation computation can be efficiently implemented using a version of EM algorithm. The proposed method is applied to handle the real data problem in 2017 Korean Household Income and Expenditure Survey (KHIES) conducted by Statistics Korea.

EO358 Room Jessel INSTABILITIES IN MULTIVARIATE DATA Chair: Zuzana Praskova

E0891: **Tests of independence of functional observations**
Presenter: **Zdenek Hlavka**, Charles University, Czech Republic

Co-authors: Marie Huskova, Simos Meintanis

Statistical problems are often simplified by assuming independence of observations. Therefore, in real life applications, one should be able to test the validity of this assumption. We investigate general tests of independence in the framework of functional data, i.e., we test the null hypothesis that the observed random curves are independent. We note that existing procedures usually test only for lack of covariance, rather than independence and, for this reason, it makes sense to propose a new procedure based on characteristic functions (CF) that should be consistent against arbitrary deviations from the null hypothesis. After establishing basic asymptotic properties of the proposed CF-based test statistic, we discuss computational issues and investigate small sample properties of the CF-based test in a simulation study considering, e.g., functional autoregression, ARCH, and GARCH alternatives.

E0278: **Automatic detection of successive variable groups in a multivariate regression model**

Presenter: **Matus Maciak**, Charles University, Czech Republic

A multivariate regression model is considered where the explanatory variables are naturally grouped into a series of consecutive groups. The number of variables and the number of groups can both increase with the sample size. However, most of the successive groups are assumed to be identical in terms of having the same effect on the response variable. The model "instabilities" occur in situations where two neighboring groups play a different role—thus, there is a sudden change with respect to the magnitude of the group effect. The idea of the proposed method is to perform an automatic detection of the different successive groups and, simultaneously, to quantify the magnitudes of the corresponding effects. The quantile check function and the fused type penalty are both combined together to obtain a robust estimate with the desired qualities. Some theoretical results are derived and the finite sample performance is investigated using a simulation study. The practical applicability of the model is illustrated with some real data examples.

E0531: **Bootstrapping in large panels with cross sectional dependence**

Presenter: **Jan Ditzen**, Heriot-Watt University, United Kingdom

Bootstrapping standard errors and confidence intervals is standard in applied econometrics. However, the literature is missing applications for large panels with heterogeneous slopes. Large dynamic panels combine characteristics of time series and panel data. In addition, large panels can contain unobserved dependence across cross sectional units, often contained in the error term of a regression model. A bootstrap has to maintain the structure of the model across the time and the cross sectional dimension and of the error term. A wild bootstrap is proposed to maintain the error structure. Results of the bootstrap with residuals in- and excluding cross sectional dependence are compared. Bootstrap methods for a static and a dynamic model are explained and compared.

E1028: **Testing structural breaks in large dynamic models based on extremal distribution**

Presenter: **Zuzana Praskova**, Charles University, Czech Republic

A linear dynamic panel data model is considered. Its coefficients can change in a time that is unknown and same for all the panels. An Erdos - Darling type test procedure to detect a change, based on the quadratic form of cumulative sums of weighted LSE residuals, is proposed. The asymptotic distribution of the test statistic as the number of panels and number of observations converge to infinity is studied, both under the null hypothesis of no change and under alternatives of a change. The finite sample behaviour of the test is also studied.

EO344 Room MAL 152 ALGEBRAIC STATISTICS Chair: Elisa Perrone

E1258: **Moment identifiability of homoscedastic Gaussian mixtures**

Presenter: **Carlos Amendola**, Technical University of Munich, Germany

Co-authors: Daniele Agostini, Kristian Ranestad

The focus is on the problem of identifying a mixture of Gaussian distributions with same unknown covariance matrix by their sequence of moments up to certain order. The approach rests on studying the moment varieties obtained by taking special secants to the Gaussian moment varieties, defined by their natural polynomial parametrization in terms of the model parameters. When the order of the moments is at most three, we prove an analogue of the Alexander-Hirschowitz theorem classifying all cases of homoscedastic Gaussian mixtures that produce defective moment varieties. As a consequence, we determine identifiability when the number of mixed distributions is smaller than the dimension of the space. In the two component setting we provide a closed form solution for parameter recovery based on moments up to order four, while in the one dimensional case we interpret the rank estimation problem in terms of secant varieties of rational normal curves.

E1372: **Total positivity in structured binary distributions**

Presenter: **Piotr Zwiernik**, Universitat Pompeu Fabra, Spain

Co-authors: Caroline Uhler, Steffen Lauritzen

Binary distributions are studied which are multivariate totally positive of order 2 (MTP2). Binary distributions can be represented as an exponential family and we show that MTP2 exponential families are convex. Moreover, MTP2 quadratic exponential families, which contain ferromagnetic Ising models and attractive Gaussian graphical models, are defined by intersecting the space of canonical parameters with a polyhedral cone whose faces correspond to conditional independence relations. Hence MTP2 serves as an implicit regularizer for quadratic exponential families and leads to sparsity in the estimated graphical model. The analysis of data from two psychological disorders will be provided.

E1080: **Supermodular inequalities in hidden variable models**

Presenter: **Anna Seigal**, University of Oxford, United Kingdom

Co-authors: Guido Montufar

The implicit semi-algebraic description of a statistical model gives a membership test based on the signs of polynomials. We discuss supermodular inequalities, which take the form of signs of conditional independence statements. We focus on two graphical models with hidden variables, both on three binary observed variables. The semi-algebraic description of the models is given in terms of supermodular inequalities. We use this description to obtain a closed form expression for the maximum likelihood estimates, and discuss supermodular inequalities of larger models.

E1530: **Exact solutions in log-concave maximum likelihood estimation**

Presenter: **Kaie Kubjas**, Aalto University, Finland

Co-authors: Alexandros Grosdos, Olga Kuznetsova, Georgy Scholten, Miruna-Stefana Sorea, Bernd Sturmfels

In nonparametric statistics one abandons the requirement that a probability density function belongs to a statistical model with finitely many parameters, and instead requires that it satisfies certain constraints. The logarithm of a probability density function is concave. The logarithm of the maximum likelihood estimate has been shown to be a piecewise linear function. We study exact solutions to log-concave maximum likelihood estimation in special cases.

EO078 Room MAL 254 TOPICS IN SPATIAL AND SPACE-TIME STATISTICS Chair: Federico Crudu

E0477: **Modeling the reduction of sample size for spatial datasets**

Presenter: **Ronny Vallejos**, Universidad Tecnica Federico Santa Maria, Chile

Co-authors: Werner Creixell, Jonathan Acosta, Javier Perez

The focus is on the reduction of sample sizes due to the effect of autocorrelation for spatial/time models. This type of problem is quite common in several disciplines where the goal is to reduce the number of georeferenced observations to be sampled. The effective sample size (ESS) is defined

as the number of independent and identically distributed observations of a spatial process. Recent research extended that definition to more complex models with more general conditions for the theoretical developments. We motivate the study of ESS with several practical problems illustrated with real-world datasets. Then, an overview of the main proposals will be presented, including a general definition of ESS for arbitrary spatial regression processes as a weighted version of the ESS of each column in the design matrix. Examples of different patterned correlation structures are explored in order to establish theoretical properties that hold for the ESS in the original case (constant mean). We study the asymptotic properties of the ML estimates of the ESS for an increasing domain framework. The consistency and asymptotic normality are established under very precise conditions. In addition, a model-free definition of the ESS is provided in order to estimate the number of observations that are necessary to plot the codispersion map (a new tool to visualize the spatial correlation) on the plane, without excessively increasing the computational time.

E0504: **Blockwise Euclidean likelihood for estimation of space-time covariance model using OpenCL in GPGPUs**
Presenter: **Victor Morales-Onate**, Facultad Latinoamericana de Ciencias Sociales, Ecuador
Co-authors: Federico Crudu, Moreno Bevilacqua

A spacetime blockwise Euclidean likelihood method is proposed for the estimation of covariance model when dealing with large spacetime Gaussian data. The method uses moment conditions coming from the score of the pairwise composite likelihood. A feature of this approach is that it is possible to obtain computational benefits with respect to pairwise likelihood depending on the choice of the spatiotemporal blocks. We also study the asymptotic properties of the proposed estimator. In order to speed-up computation we consider a general purpose GPU implementation using OpenCL. We illustrate the advantages of our methodology by means of a simulation study highlighting the computational gains of the OpenCL GPU implementation. Finally, we apply our estimation method to the Irish wind speed data.

E0840: **Estimation of the structural similarity index for remote sensing data**
Presenter: **Felipe Osorio**, Universidad Tecnica Federico Santa Maria, Chile
Co-authors: Ronny Vallejos, Wilson Barraza, Silvia Ojeda, Marcos Landi

The structural similarity (SSIM) index has been studied from different perspectives in the last decade. Most of the developments consider its parameters fixed. Because each of these parameters corresponds to the weight of a factor in the final SSIM coefficient, the usual assumption that all parameters are equal to one is questionable. A new model-based estimation method is proposed and developed so that, the usual assumption that all parameters are equal to one can be handled via approximate hypothesis-testing techniques that are properly developed in the context of regression. The method considers nonlinear models with multiplicative noise to explain the root mean square error (RMSE) as a function of the SSIM index for two given images that are split into several subimages to generate the samples necessary for the regression models. The nonlinear model is estimated using a pseudo-likelihood approach for which a recursive estimation algorithm is provided. A numerical experiment based on a Monte Carlo simulation is provided to test whether the parameters are all equal to one and to gain more insight into the performance of the estimates in practice.

E1136: **Spatial models in the space of covariates: Methodological and computational issues**
Presenter: **Flavio Santi**, University of Verona, Italy
Co-authors: Maria Michela Dickson, Diego Giuliani, Giuseppe Espa

Spatial dependence in the space of covariates of a regression model may arise for several reasons, including spatial trends of covariates, and model misspecifications. Just like spatial dependence over geographical or physical spaces, dependence of regression residuals in the space of covariates may lead to inconsistent estimates of regression parameters as well as of standard errors, thereby making inference unreliable. Although standard methodologies of spatial econometrics also hold when modelling dependence in the space of covariates, the very nature of that space poses new methodological and computational issues, mainly because of the dimensionality of the space of covariates, and because of the lack of isotropy in the spatial dependence. Both problems are analysed both from a theoretical and a computational point of view. In particular, it is studied how anisotropies should be modelled when the space of covariates is used as a proxy of the geographical space.

EO130 Room Senate STATISTICS OF SPATIOTEMPORAL EXTREMES Chair: Raphael Huser

E0812: **Approximate Bayesian inference for spatial flood frequency analysis**
Presenter: **Birgir Hrafnkelsson**, University of Iceland, Iceland
Co-authors: Raphael Huser, Arni Johannesson, Haakon Bakka, Stefan Siegert

Extreme floods cause casualties and damage to vital civil infrastructure. Predictions of extreme floods within gauged and ungauged catchments are crucial to mitigate these disasters. A latent Gaussian model is proposed for predicting extreme floods using the generalized extreme-value (GEV) distribution and a novel multivariate link function for its location, scale and shape parameters. This link function is designed to separate the interpretation of the parameters at the latent level and to avoid unreasonable estimates of the shape parameter. Structured additive regression models are proposed for the three parameters at the latent level. Each of these regression models contains fixed linear effects for catchment descriptors. Spatial model components are added to the two first latent regression models, to model the residual spatial structure unexplained by the catchment descriptors. To achieve computational efficiency for large datasets with these richly parametrized models, we exploit a Gaussian-based approximation to the posterior density. This approximation relies on site-wise estimates, but, contrary to typical plug-in approaches, the uncertainty in these initial estimates is properly propagated through to the final posterior computations. We applied the proposed modeling and inference framework to annual peak river flow data from 554 catchments across the United Kingdom. The framework performed well in terms of flood predictions for ungauged catchments.

E0881: **Spatio-temporal clustering of extremal behaviour for environmental variables**
Presenter: **Christian Rohrbeck**, Lancaster University, United Kingdom

To address the need for efficient inference for a range of extreme value problems, spatial pooling of information is the standard approach for marginal tail estimation. A spatial-temporal clustering method is introduced which accounts for both the similarity of the marginal tails and the spatial-temporal dependence structure of the data to determine the appropriate level of pooling. Spatio-temporal dependence is incorporated in two ways: to determine the cluster selection and to account for dependence of the data over sites within a cluster when making the marginal inference. We propose a statistical model for the pairwise extremal dependence which accounts for distance across space and time, and accommodates our belief that sites within the same cluster tend to exhibit a higher degree of dependence than sites in different clusters. We use a Bayesian framework which learns about both the number of clusters and their spatial-temporal structure, and that enables the inference of site-specific marginal distributions of extremes to incorporate uncertainty in the clustering allocation. The approach is illustrated based on the analysis of daily river flow levels in the UK.

E0610: **Spatio-temporal extremes based on single-location conditioning**
Presenter: **Jenny Wadsworth**, Lancaster University, United Kingdom
Co-authors: Emma Simpson

Conditional approaches for modelling multivariate extreme events have benefits in terms of computational efficiency and their ability to capture both asymptotic dependence and asymptotic independence. We present such a model for the spatio-temporal setting, conditioning on the occurrence of an extreme value at single location in space, and a single point in time. Combining information across all conditioning sites allows for inference

via a composite likelihood approach. We apply our model to Red Sea surface temperature data, and present and discuss diagnostic approaches to assess the model fit.

E1026: **A semiparametric Bayesian model for large spatiotemporal Red sea surface temperature data and related hotspot estimation**
Presenter: **Arnab Hazra**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser

Gradually increasing sea surface temperature (SST) is a major concern for the ecosystems and we focus on pointing out the exceedance regions of SST within the Red Sea, a vital region of endangered coral reefs. Explosive growth of remote sensing and other data collection techniques are leading to large high resolution spatial datasets. We propose a Dirichlet process mixture of low-rank spatial Student-t processes for spatial analysis of large datasets where temporal replications (days, for example) are available. Our model considers the whole dataset above a very low threshold for modeling the bulk and the tail jointly. The model allows drawing extremal inference by probabilistically separating the extreme days from the moderate days and estimating the model parameters based on the cluster of extreme days. The proposed model has nonstationary mean, covariance and asymptotic dependence and under limiting conditions, it spans all possible spatial processes. Inference is drawn based on MCMC sampling where most of the parameters allow Gibbs sampling. We perform a simulation study to compare model fitting performances of the proposed model with its sub-models where our method generally outperforms its alternatives. Finally, we fit the proposed model to estimate the spatial return level maps and to identify the exceedance regions. Compared to the low-rank Gaussian processes, estimated return levels based on the proposed model are generally higher across the Red Sea.

EO098 Room CLO 203 MODERN METHODS IN THE ANALYSIS OF DIRECTIONAL DATA Chair: Jose Ameijeiras-Alonso

E0607: **Copula-based segmentation of cylindrical time series**
Presenter: **Francesco Lagona**, University Roma Tre, Italy

Bivariate sequences of angles and intensities are often referred to as cylindrical time series, because the pair of an angle and an intensity can be represented as a point on a cylinder. In environmental studies, examples of these data include time series of wind directions and pollutant concentrations, wind directions and speeds and wave directions and heights. The analysis of cylindrical time series is complicated by the difficulties in modeling the dependence between angular and linear measurements and the temporal correlation between cylindrical observations. An additional complication that often arises in environmental studies is the multimodality of the marginal distribution of the data because environmental cylindrical data are typically observed under heterogeneous conditions that vary over time. A parsimonious hidden Markov model is introduced to simultaneously account for linear-circular dependence, temporal auto-correlation and multimodality. Under this model, the distribution of cylindrical data is approximated by a mixture of copula-based cylindrical densities, whose parameters depend on the evolution of a latent Markov chain. While the copula-based cylindrical density accommodates linear-circular dependence, a mixture of copula-based densities allows for multimodality and, finally, a latent Markov chain accounts for temporal correlation.

E0798: **Nonparametric circular regression estimation with spatially correlated errors**
Presenter: **Andrea Meilan-Vila**, Universidade da Coruna, Spain
Co-authors: Mario Francisco-Fernandez, Rosa Crujeiras, Agnese Panzera

Circular data can be regarded as points whose support is on a circle (with unit radius) measured in degrees or radians and with periodic nature. Examples of circular data arise in many applied fields such as biology (animal orientation), meteorology (wind direction) or oceanography (ocean currents), among others. These data may exhibit an important feature: close observations tend to be more similar than those that are far apart. Therefore, such observations cannot be treated as independent and the dependence structure should be taken into account in the estimation process. The aim is to propose and study nonparametric procedures to estimate the circular regression function, assuming a multivariate linear-circular regression model (circular responses and multivariate linear predictors) with spatially correlated circular errors. The new approaches consist in computing the inverse tangent function of the ratio between kernel estimators of the conditional expectation of the sine and cosine of the response, respectively. Nadaraya-Watson and local polynomial type estimators are considered. The asymptotic bias and variance of the proposed nonparametric estimators are derived. Additionally, some guidelines to select asymptotically local optimal matrix bandwidths are given. Simulation studies are carried out to check the finite sample performance of the considered estimators. The methodology is illustrated with a real data set.

E1202: **The hyper-flexible family of distributions on the circle**
Presenter: **Domien Craens**, UGent, Belgium
Co-authors: Christophe Ley

A new family is proposed for circular distributions that is highly versatile in the sense that it is able to model any data structure, symmetric or asymmetric, peaked or flat-topped, uni- or multi-modal. Therefore, we coin the term hyper-flexible family of distributions. It is based on a finite sum of trigonometric functions, and we derive various properties such as the cumulative distribution function, trigonometric moments and random number generation. We discuss in detail parameter estimation via maximum likelihood, and provide both theoretical and practical considerations. A thorough Monte Carlo simulation study shows the superior flexibility of our new family as compared to existing models from the literature. We also analyze three notoriously complicated data sets with our hyper-flexible family of distributions.

E1223: **Wrapped normal graphical models**
Presenter: **Anna Gottard**, University of Firenze, Italy
Co-authors: Agnese Panzera

Directional distributions are widely used in many research fields such as biology, medicine, geography or meteorology. Most of the studies focus on the univariate or bivariate case, but some interesting fields require higher dimensional settings. In this framework, the protein structure prediction problem is considered of great interest. Graphical models are a widely used class of multivariate models, where the conditional independence structure of a set of variables can be summarised by a graph. The random variables are represented by the nodes in the graph and conditional independence by missing edges. We introduce a general theory for graphical models for the multivariate Wrapped Normal distribution that allows studying the conditional independence structure of the dihedral angles of each amino acid of a protein. Wrapped Normal graphical models inherit most of the properties of the ordinary Gaussian graphical models, such as closeness to conditioning and marginalisation, decomposability and so on. We provide an interesting interpretation of model parameters and discuss inferential issues.

EO500 Room CLO 204 BAYESIAN APPROACHES TO THE ANALYSIS OF NEUROIMAGING Chair: John Kornak

E0333: **Bayesian generalized sparse symmetric tensor-on-vector regression**
Presenter: **Rajarshi Guhaniyogi**, University of California Santa Cruz, United States

Motivated by brain connectome datasets acquired from various imaging modalities, a novel generalized Bayesian linear modeling framework with a symmetric tensor response and scalar predictors is proposed. The symmetric tensor coefficients corresponding to the scalar predictors are embedded with two features: low-rankness and group sparsity within the low-rank structure. Besides offering computational efficiency and parsimony, these two features enable identification of important tensor nodes and tensor cells significantly associated with the predictors. We establish that the posterior predictive density from the proposed model is close to the true density, the closeness being measured by the Hellinger distance between these two densities, which scales at a rate nearing the finite dimensional optimal rate of square root of the sample size, depending on how the

number of tensor nodes grows with the sample size. The proposed framework is empirically investigated under various simulation settings and with a brain connectome dataset.

E0534: **Spatial 3D Matern priors for fast whole-brain fMRI analysis**
Presenter: **Per Siden**, Linkoping University, Sweden
Co-authors: Finn Lindgren, David Bolin, Anders Eklund, Mattias Villani

Bayesian whole-brain functional magnetic resonance imaging (fMRI) analysis with three-dimensional spatial smoothing priors have been shown to produce state-of-the-art activity maps without pre-smoothing the data. The proposed inference algorithms are computationally demanding however, and the proposed spatial priors have several less appealing properties, such as being improper and having infinite spatial range. We propose a statistical inference framework for functional magnetic resonance imaging (fMRI) analysis based on the class of Matern covariance functions. The framework uses the Gaussian Markov random field (GMRF) representation of Matern fields via the stochastic partial differential equation (SPDE) approach. This allows for more flexible and interpretable spatial priors, while maintaining the sparsity required for fast inference in the high-dimensional whole-brain setting. We develop an accelerated stochastic gradient descent (SGD) optimisation algorithm for empirical Bayes (EB) inference of the spatial hyperparameters. Conditional on the inferred hyperparameters, we make a fully Bayesian treatment of the main parameters of interest, that is, the brain activity coefficients. We apply the Matern prior to both experimental and simulated task-fMRI data and clearly demonstrate that this is a more reasonable choice than the previously used priors, by using prior simulation, cross validation and visual inspection of the resulting activation maps.

E1562: **Bayesian approaches for dynamic brain connectivity**
Presenter: **Michele Guindani**, University of California, Irvine, United States

A Bayesian framework for estimating time-varying functional connectivity networks from brain fMRI data will be discussed. Dynamic functional connectivity, i.e., the study of how interactions among brain regions change dynamically over the course of a fMRI experiment, has recently received wide interest in the neuroimaging literature. The method utilizes state space models for classification of latent neurological states, achieving estimation of the connectivity networks in an integrated framework that borrows strength over the entire time course of the experiment. Furthermore, we assume that the graph structures, which define the connectivity states at each time point, are related within a super-graph, to encourage the selection of the same edges among related graphs.

E1977: **Covariate-adjusted hybrid principal components analysis for EEG data**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States
Co-authors: Abigail Dickinson, Charlotte DiStefano, Shafali Jeste, Damla Senturk

Electroencephalography (EEG) studies produce region-referenced functional data in the form of signals recorded across electrodes on the scalp. The data capture underlying neural dynamics, and it is of clinical interest to model differences in neurodevelopmental trajectories between diagnostic groups, e.g. typically developing (TD) children and children with autism spectrum disorder (ASD). Valid group level inference requires characterization of the EEG dependency structure and covariate-dependent heteroscedasticity, such as changes in variation over developmental age. Resting state EEG is collected on both TD and ASD children aged two to twelve years old. The peak alpha frequency (PAF) is an important biomarker linked to neurodevelopment. It is known to shift from lower to higher frequencies as children age. We model patterns of alpha spectral variation, rather than just the peak location, regionally across the scalp and chronologically across development for both the TD and ASD diagnostic groups. We propose a covariate-adjusted hybrid PCA (CA-HPCA) for region-referenced functional EEG data. It utilizes both vector and functional PCA while simultaneously adjusting for covariate-dependent heteroscedasticity. CA-HPCA assumes the covariance process is weakly separable conditional on observed covariates leading to stable and computationally efficient estimation. A mixed effects estimation framework is proposed coupled with a bootstrap test for group level inference.

EO534 Room MAL 251 STATISTICS OF RANDOM PROCESSES FOR ANALYSING HIGH FREQUENCY DATA Chair: Masayuki Uchida

E0585: **Statistics for stochastic PDEs based on high-frequency observations**
Presenter: **Markus Bibinger**, Philipps University of Marburg, Germany
Co-authors: Mathias Trabs

Parameter estimation is discussed for a parabolic, linear stochastic partial differential equation (SPDE) from observations of a solution on a discrete grid in time and space. We consider SPDEs with one spatial dimension and a bounded spatial domain with Dirichlet boundary conditions. Focusing first on volatility estimation and assuming a high-frequency regime in time, we provide an explicit and easy to implement method of moments estimator based on squared time increments. Our estimator is consistent and admits a central limit theorem. The asymptotic theory is developed based on a representation of the solution of the SPDE as an infinite SDE-factor model and exploiting central limit theorems for time series which satisfy some mixing-type properties. This is established moreover for the joint estimation of the integrated volatility and parameters in the differential operator in a semi-parametric framework.

E0673: **Noise estimation for ergodic Levy driven SDE in YUIMA package**
Presenter: **Hiroki Masuda**, Kyushu University, Japan
Co-authors: Yuma Uehara, Lorenzo Mercuri

Levy driven stochastic differential equation (SDE) is a flexible building block for modeling non-Gaussian high-frequency data observed in many application fields such as biology and ecology. It is, however, common knowledge that a closed form of the likelihood function is rarely available except for quite special cases, making estimation of characteristics of the driving Levy noise difficult. We propose a multistep estimation procedure, by utilizing the Euler residuals constructed from the Gaussian quasi-maximum likelihood estimator (GQMLE); specifically, we first estimate the parametric coefficient by the GQMLE, next approximate unit time increments of the driving noise by partially summing up the Euler residuals, and then apply M-estimation theory (parametric or not). We will present large-sample properties of the proposed estimator, followed by numerical experiments through the YUIMA package in R.

E1499: **Reguralized estimation with multiple penalties and its application to stochastic differential equations**
Presenter: **Alessandro De Gregorio**, University of Rome La Sapienza, Italy
Co-authors: Francesco Iafrate

Penalized estimation methods for statistical models with different rates of convergence are introduced. Furthermore, we discuss the asymptotic properties of our estimator. In particular, we focus our attention on stochastic differential equations observed at discrete times. Some numerical examples are also provided.

E1431: **Estimation for degenerate diffusion processes**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan
Co-authors: Arnaud Gloter

A multi-dimensional ergodic diffusion process specified by a system of stochastic differential equations is considered. The first component has a non-degenerate matrix diffusion coefficient and the second component has no diffusion coefficient. Each coefficient has an unknown vector parameter and we estimate these parameters based on long-term high frequency observations. Adaptive and non-adaptive methods are discussed. The convergence rates of the diffusion parameter and the drift parameter in the non-degenerate component are the same as the usual ones but the

asymptotic variance is improved. The convergence of the estimator for the parameter in the degenerate component is much faster than others.

EO494 Room MAL 252 ADVANCES IN CLASSIFICATION AND HIGH DIMENSIONAL STATISTICS Chair: Daoji Li

E0204: **Functional partial linear quantile regression based on reproducing kernel Hilbert space**
Presenter: **Peng Liu**, University of Kent, United Kingdom
Co-authors: Nan Zhang, Bei Jiang, Linglong Kong, Jianhua Huang

Functional and nonfunctional data are often encountered simultaneously in modern experiments for example the clinical trial as well in economics. However, it is difficult to consider both data at the same time. We consider functional partial linear quantile regression, where both infinite dimension functional as well as finite dimension slope parameters are included. We study the theoretical properties under a reproducing kernel Hilbert space framework which was being proved to be very flexible and powerful. Under this framework, we also developed an ADMM algorithm which is very easy to implement in practical applications. Simulation studies and real data studies are performed to validate our propose methodology and practical applications respectively.

E0229: **Classification with imperfect training labels**
Presenter: **Timothy Cannings**, University of Edinburgh, United Kingdom
Co-authors: Yingying Fan, Richard Samworth

The effect of imperfect training data labels on the performance of classification methods is studied. In a general setting, where the probability that an observation in the training dataset is mislabelled may depend on both the feature vector and the true label, we bound the excess risk of an arbitrary classifier trained with imperfect labels in terms of its excess risk for predicting a noisy label. This reveals conditions under which a classifier trained with imperfect labels remains consistent for classifying uncorrupted test data points. Furthermore, under stronger conditions, we derive detailed asymptotic properties for the popular k-nearest neighbour ($k$nn), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. One consequence of these results is that the $k$nn and SVM classifiers are robust to imperfect training labels, in the sense that the rate of convergence of the excess risks of these classifiers remains unchanged. On the other hand, the LDA classifier is shown to be typically inconsistent in the presence of label noise unless the prior probabilities of each class are equal.

E1131: **Simultaneous prediction intervals for high-dimensional vector autoregressive model**
Presenter: **Mengyu Xu**, University of Central Florida, United States
Co-authors: Sayar Karmakar

The simultaneous prediction intervals for high-dimensional vector autoregressive model are studied. We consider a de-biased calibration for the lasso prediction and propose a Gaussian-multiplier bootstrap based method for one-step ahead prediction. The asymptotic coverage consistency of the prediction interval is obtained. We also develop simulation results to evaluate the finite sample performance of the procedure.

E1314: **Bayesian variable selection for joint mean and covariance models**
Presenter: **Jiaming Shen**, The University of Manchester, United Kingdom
Co-authors: Jianxin Pan

Modelling covariance matrices is generally difficult due to two obstacles, i.e., high-dimensionality and positive definiteness. Based on a modified Cholesky decomposition (MCD) of the covariance matrix, we propose to model the mean, the generalised autoregressive parameters and the innovation variances resulting from the MCD, simultaneously, in terms of linear regression models. We consider not only the parameter estimation, but also variable selection through Bayesian analysis. Specifically, Markov chain Monte Carlo (MCMC) sampling strategy is considered and Gibbs sampler is used to draw random samples from the posterior distributions of the model parameters. Bayesian variable selection methods through adding certain shrinkage penalty terms are also investigated. A newly developed R package called BayesJMCM is introduced and its use is demonstrated through both simulated data and real data. A comparison to existing frequency methods is made, showing that more uncertainties from different sources of the models are accounted and more reliable results are obtained.

EO336 Room MAL 253 NEW METHODS FOR MODELLING ORDINAL AND MIXED-TYPE DATA Chair: Cristina Mollica

E0468: **Mixture of Hidden Markov models for accelerometer data**
Presenter: **Marie du Roy de Chaumaray**, CREST-ENSAI, France
Co-authors: Matthieu Marbac, Fabien Navarro

The motivation comes from the analysis of accelerometer data. The analysis of such data consists in extracting statistics which characterize the physical activity of a subject (*e.g.,* the mean time spent at different activity levels and the probability of the transition between two levels). Therefore, we introduce a finite mixture model of hidden Markov chain to analyze accelerometer data by considering heterogeneity into the population. This approach does not specify activity levels in advance but estimates them from the data. In addition, it allows for the heterogeneity of the population to be taken into account and defines subpopulations having a homogeneous behavior regarding the physical activity. The main theoretical result is that, under mild assumptions, the probability of misclassifying an observation decreases at an exponential rate with its length. Moreover, we prove the model identifiability, and we show how the model can handle missing values. Our proposition is illustrated using real data.

E0580: **Bayesian rank aggregation for (mixtures of) Plackett-Luce models**
Presenter: **Stephen Johnson**, Newcastle University, United Kingdom

Ranked data are central to many applications in science and social science and arise when rankers (individuals) use some criterion to order a set of entities. Rank aggregation aims to produce a single overall ranking that is representative of a collection of rankers. One approach is to consider models that rely on strong assumptions of homogeneity. However, in general, this assumption is not likely to be plausible. We suggest that the data should instead be modelled in a manner as flexible as possible; with the intention of obtaining good model fit. Then, once an adequate model has been obtained, the aggregate ranking should be that with the largest predictive probability, that is, the mode of the posterior predictive distribution. Unfortunately, the dimension of the predictive distribution grows factorially with the number of entities, so it is often unobtainable. We consider methods for performing posterior predictive checks and also for obtaining the aggregate ranking under (mixtures of) Plackett-Luce models; with approximations becoming necessary when the number of entities is large. The methodology is illustrated through simulation studies and we provide insight as to when approximations are likely to perform well.

E0723: **Residuals diagnostics for the choice of model-based trees for ordinal responses**
Presenter: **Rosaria Simone**, University of Naples Federico II, Italy

Model-based profiling of preferences and evaluations, when collected as either rating or marginal ranking data, can be successfully pursued by means of classification and regression trees. In this regard, the model-based approach should adopt flexible and parsimonious models for the tree nodes in order to enable the understanding of the rating process and the derivation of response profiles. CUBREMOT is a class of model-based binary trees for preference and evaluation data that is grounded on the specification of CUB models, of which the Binomial is a particular case. The flexibility of CUBREMOT in shaping rating data at different subsetting levels, also in presence of some structural inflated categories, is enhanced by diagnostic checks in terms of residuals. For ordinal responses, the analysis of residuals built via a jittering approach to perform model selection can be useful to assess the goodness of the derived classification, to choose the best baseline model, to identify the best splitting criterion (if more

are available), or to tune the tree depth for the post-pruning phase. The proposal is discussed on the basis of several data examples to show its efficacy and the extent of its applicability.

E0788: **Comparison between likelihood-based methods of factor models for ordinal data**
Presenter: **Silvia Cagnone**, University of Bologna, Italy
Co-authors: Silvia Bianconcini

Latent variable models represent a useful tool in different fields of research in which the constructs of interest are not directly observable. In presence of many latent variables/random effects, problems related to the integration of the likelihood function can arise since analytical solutions do not exist. In literature, different solutions have been proposed to overcome these problems. Among these, the composite likelihoods method and more recently the dimension-wise method have been shown to produce estimators with desirable properties. We compare the performance of the two methods in the case of longitudinal ordinal data.

EO568 Room SH349 MODERN APPROACHES TO THE SPECTRAL ANALYSIS OF TIME SERIES Chair: Tobias Kley

E0335: **Nonlinear spectral analysis: A local Gaussian approach**
Presenter: **Lars Arne Jordanger**, Western Norway University of Applied Sciences, Norway

The spectral distribution $f(\omega)$ can detect periodicites in a stationary time series Y_t, but it has some limitations due to its dependence on the autocorrelations $\rho(h)$. $f(\omega)$ completely determines Gaussian time series, but it is an inadequate tool when Y_t contains asymmetries and nonlinear dependencies (it can e.g. not distinguish white i.i.d. noise from GARCH-type models, whose terms are dependent, but uncorrelated). A local Gaussian spectral distribution $f_v(\omega)$ enables a local investigation of Y_t by replacing the autocorrelations $\rho(h)$ with local Gaussian autocorrelations $\rho_v(h)$. A key feature of $f_v(\omega)$ is that it coincides with $f(\omega)$ for Gaussian time series, which implies that $f_v(\omega)$ can be used to detect non-Gaussian traits in other time series. If $f(\omega)$ is flat, then peaks and troughs of $f_v(\omega)$ can indicate nonlinear traits, which potentially might discover local periodic phenomena that goes undetected in an ordinary spectral analysis.

E0379: **Extending the validity of frequency domain bootstrap methods to general stationary processes**
Presenter: **Marco Meyer**, TU Braunschweig, Germany
Co-authors: Efstathios Paparoditis, Jens-Peter Kreiss

Existing frequency domain methods for bootstrapping time series have a limited range. Essentially, these procedures cover the case of linear time series with independent innovations, and some even require the time series to be Gaussian. We propose a new frequency domain bootstrap method – the hybrid periodogram bootstrap (HPB) – which is consistent for a much wider range of stationary, even nonlinear, processes and which can be applied to a large class of periodogram-based statistics. The HPB is designed to combine desirable features of different frequency domain techniques while overcoming their respective limitations. It is capable to imitate the weak dependence structure of the periodogram by invoking the concept of convolved subsampling in a novel way that is tailor-made for periodograms. We show consistency for the HPB procedure for a general class of stationary time series, ranging clearly beyond linear processes, and for spectral means and ratio statistics, on which we mainly focus. The finite sample performance of the new bootstrap procedure is illustrated via simulations.

E0737: **An exact expression for the Whittle approximation of the Gaussian likelihood and its application to parameter estimation**
Presenter: **Suhasini Subbarao**, Texas A&M, United States

One important objective in time series analysis is to estimate the parameters of a conjectured time series model based on the observed time series. If the parameters in the model can be characterized in terms of their second order autocovariance structure, then there are two well-known methods for estimating the parameters; one in the time domain, the other in the frequency domain. The time domain approach is based on the (quasi) Gaussian likelihood, whereas in the frequency domain the Whittle likelihood is commonly used. The Whittle likelihood can be viewed as the Kullbach-Leibler distance between the periodogram and the conjectured spectral density. It is well known that the Whittle likelihood is an approximation of the Gaussian likelihood. We obtain an exact expression for this approximation in terms of biorthogonal random variables. This approximation can be used to obtain an alternative proof of Szego's theorem. We apply this approximation to the problem of parameter estimation, where we obtain a variant of the Whittle likelihood which has better finite sample properties.

E1238: **Asymptotic normality of integrated periodogram operators**
Presenter: **Daniel Rademacher**, TU Braunschweig, Germany
Co-authors: Jens-Peter Kreiss, Efstathios Paparoditis

Consider a strictly stationary functional process. A key element of a frequency domain framework for drawing statistical inference on the second-order structure of the process is the spectral density operator, which generalises the notion of a spectral density matrix to the functional setting. As an integral operator, the spectral density operator is fully determined by its corresponding kernel, which can be estimated by a smoothed version of the periodogram kernel (the functional analogue to the periodogram matrix). More generally, many interesting quantities of the process such as autocovariance operators, or the spectral distribution operator can be represented as a weighted integral of the spectral density kernel. Estimators for such a quantity are obtained by exchanging the spectral density kernel with the periodogram kernel. Thus, the class of integrated periodogram operators covers many familiar statistics, including empirical autocovariance and smoothed periodogram operators. We show that any finite collection of such estimators converges to a collection of jointly complex normal distributed operators. As a side-result, we obtain the joint asymptotic normality for the empirical autocovariance operators. These results do not depend on any structural modelling assumptions, but only on functional versions of cumulant mixing conditions.

EC809 Room MAL 354 CONTRIBUTIONS IN ROBUST STATISTICS Chair: Agustin Mayo-Iscar

E1613: **Robust Bayesian estimation by using the quasi-posterior with divergence**
Presenter: **Tomoyuki Nakagawa**, Tokyo University of Science, Japan

In Bayesian analysis, it is well known that ordinary Bayesian estimators are not robust against outliers. Recently, robust Bayesian estimation against outliers has been proposed by using quasi-posterior and robust divergences. There are two type robust Bayesian estimators, one using the density power divergence and one using the γ-divergence. The robustness is characterized in term of the influence function. However, the calculation of the influence function is not easy. Furthermore, the estimator using the density power divergence does not work well for the estimation of the scale parameter, and it is unstable when the contamination ratio is not small. These properties are same from the frequentist viewpoint. On the other hand, from the frequentist viewpoint, it is well known that an estimator using the γ-divergence can make an estimation stable even when the contamination ratio is not small. Thus, we focus on the estimation using the gamma-divergence and compare the two type robust Bayesian estimations. We show the performance of this robust Bayes estimation in various situations.

E1890: **Robustness and shrinkage for GLM with the forward search**
Presenter: **Fabrizio Laurini**, University of Parma, Italy

Supervised methods of classification naturally exploit linear and non linear relationships between explanatory variables and a response. However, the presence of clusters may lead to a different pattern within each group. For instance, data can naturally be grouped in several linear structures and so, even a linear regression models can be used for classification. Estimation of linear models can be severely biased by influential observations or outliers. A practical problem arises when the groups identifying the different relationships are unknown, and the number of relevant variables

is high. In such a context, supervised classification problem can become cumbersome. As a solution, within the general framework of generalized linear models, a new robust approach is to exploit the sequential ordering of the data provided by the forward search algorithm. Such an algorithm will be used two-folds to address the problems of variable selection for model fit, while grouping the data naturally around the model. The influence of outliers, if any is inside the dataset, will be monitored at each step of the sequential procedure. Preliminary results on simulated data have highlighted the benefit of adopting the forward search algorithm, which can reveal masked outliers, influential observations and show hidden structures.

E1857: **Robust non-ordinal polytomous regression**
Presenter: **Benjamin Poilane**, University of Geneva, Switzerland
Co-authors: Julien Miron, Eva Cantoni

Data with non-ordered categorical responses occur in various fields (medicine, sociology, political science). Polytomous regression is a well-used tool to make inference on such data. However, datasets can often contain outlying covariate values or mislabelled responses. In such cases, classical maximum likelihood inference may give disproportionate influence to very few points and thus highly bias the estimation. To counter such effects, one can use robust methods. Existing robust polytomous regression estimators are reviewed, and two new proposals are introduced: a robust GLM-based estimator and the optimal self-standardized B-robust estimator with corresponding Wald-type test statistics. Asymptotic properties of these two methods are derived. Robustness properties and computational costs of the existing and new estimators are compared theoretically and through extensive simulation study. The interest of the methods is illustrated on real datasets.

E2006: **Robust variable selection for model-based learning in presence of adulteration**
Presenter: **Andrea Cappozzo**, University of Milano Bicocca, Italy
Co-authors: Francesca Greselin, Thomas Brendan Murphy

The problem of identifying the most discriminating features when performing supervised learning has been extensively investigated in the past years. In particular, several methods for variable selection in model-based classification have been proposed. Surprisingly, the impact that outliers and wrongly labelled units cause on the determination of relevant predictors has received far less attention, with almost no dedicated methodologies available in the literature. We introduce two robust variable selection approaches: one that embeds a robust classifier within a greedy-forward procedure and the other based on the theory of maximum likelihood estimation and irrelevance. The former recasts the feature identification as a model selection problem, while the latter regards the relevant subset as a model parameter to be estimated. An experiment on synthetic data is provided to underline the benefits of the proposed methods in contrast with non-robust solutions. An application to a high-dimensional classification problem of contaminated spectroscopic data is provided.

EC813 Room MAL 355 CONTRIBUTIONS IN SURVIVAL ANALYSIS Chair: Taoufik Bouezmarni

E0357: **Smoothed time-dependent ROC curve for right censored survival data**
Presenter: **Kassu Mehari Beyene**, UCLouvain, Belgium
Co-authors: Anouar El Ghouch

The prediction reliability is of primary concern in many clinical studies when the objective is to develop new predictive models. In fact, prior using a model in any clinical decision making, it is very important to check its ability to discriminate between subjects who are in risk of developing certain disease from those who will not. To that end, the time-dependent receiver operating characteristic curve (ROC) is the most commonly used method in practice. Several approaches have been proposed in the literature to estimate the ROC non-parametrically in the context of survival data. But, except one recent approach, all the existing methods provide a non-smooth ROC estimator whereas, by definition, the ROC curve is smooth. We propose and study a new non-parametric smooth ROC estimator based on a weighted kernel smoothing. As bandwidth is the main parameter to be set, we present and study different methods to appropriately select one. A simulation study is conducted, under different scenarios, to prove the consistency of the proposed method and compare its finite sample performance with a competitor. The results show that the proposed method performs better. Furthermore, we illustrate the method using a real data example.

E1710: **Higher order approximations in the Cox proportional hazards model**
Presenter: **Aneta Andrasikova**, Palacky University Olomouc, Czech Republic
Co-authors: Eva Fiserova

Time-to-event analysis can be applied in a wide range of sectors, such as medicine, economy and others. Its main idea comes from the evaluating of the time until the occurrence of an event of interest. The effect of some particular covariates on survival time can be described by the Cox proportional hazards model. The statistical significance of the effect of the considered covariates is verified by the likelihood ratio test, the Wald test, or the score test. These tests represent the first-order approximations which are asymptotically equivalent. They can lead to the numerically different results in applications according to available data. In addition to the standard test, higher-order asymptotics based on Barndorff-Nielsen and Lugannani-Rice formulas is applied for more accurate approximations. Comparison of the size, power, and adjusted power of these tests for small samples is performed on simulated datasets in dependence on the distributions of baseline hazard functions, various proportion of right censored data and the number and the distribution of the covariates.

E1966: **Nonparametric predictive inference for discrete time survival data related to the actuarial estimator**
Presenter: **Ali Mahnashi**, Durham University, United Kingdom
Co-authors: Frank Coolen, Tahani Coolen-Maturi

The hazard function is the most common representation of the event time distribution. In discrete time, the hazard at time t_j is defined as the conditional probability that a randomly selected individual will experience the event at time t_j, given that the individual did not experience the event prior to t_j. The discrete-time hazard can be estimated by the actuarial estimator of the hazard function. Nonparametric predictive inference (NPI) is a frequentist statistics method based on only few assumptions. It focuses explicitly on future observations and uses imprecise probabilities to quantify uncertainty. NPI has been presented for Bernoulli data as well as for right-censored data. We utilise the NPI lower and upper probabilities for Bernoulli data for the actuarial estimator. This development leads us to derive the NPI lower and upper survival functions for the next discrete random variable. Then we compare the discrete-time NPI lower and upper survival functions with the NPI lower and upper survival functions for right censored data in the continuous time case. Furthermore, we aim to develop the discrete-time NPI lower and upper probabilities for multiple discrete random variables. Finally, we use an illustration example to clarify our contribution.

E1597: **New insight into the role of intangible heterogeneity of covariate effects in hidden subpopulation subject to censoring**
Presenter: **Farhad Shokoohi**, University of Nevada Las Vegas, United States

The advent of modern technology has led to a surge of high-dimensional data in biology and health sciences such as genomics, epigenomics and medicine. The high-grade serous ovarian cancer (HGS-OvCa) data reported by The Cancer Genome Atlas (TCGA) Research Network is one example that includes information on over 9,000 genes. The focus is on the relationship between Disease Free Time (DFT) after surgery among ovarian cancer patients and their DNA methylation profiles of genomic features. Such studies pose additional challenges beyond the typical big data problem due to intangible population substructure and censoring. Despite the availability of several methods for analyzing time-to-event data with a large number of covariates but a small sample size, there is no method available to date that accommodates the additional feature of heterogeneity. We propose a regularized framework based on the finite mixture of accelerated failure time model to capture intangible heterogeneity due to population substructure and to account for censoring simultaneously. Our data analysis indicates the existence of heterogeneity in the HGS-OvCa

data, with one component of the mixture capturing a more aggressive form of the disease, and the second component capturing a less aggressive form. In particular, the second component portrays a significant positive relationship between methylation and DFT for BRCA1.

EG171 Room MAL 153 CONTRIBUTIONS IN CLUSTERING Chair: Pietro Coretto

E1874: **Multichannel qualitative harmonic analysis for two-step patient pathway**
Presenter: **Pierre-Louis Bithorel**, Ined, France
Co-authors: Elisabeth Morand, Gustavo De Santis

The SNIIRAM (Systeme National d Information Inter-Regimes de l Assurance Maladie) database, collected by the French National Insurance, is a very rich source of information on the drugs prescribed to patients: type, date of delivery and quantity are recorded. However, some pieces of information are missing, e.g. when drugs are actually taken and what disease is treated. A patient pathway is defined as a sequence of drugs delivered to the same person. A treatment consists of a drug, or a set of drugs, prescribed to a patient for a specific purpose (intention to cure), either as first treatment or for continuation of the treatment. A two-step patient pathway is assuming. An unsupervised multichannel qualitative harmonic analysis clustering is performed to identify patterns of first fertility treatment intention. Our approach uses a fuzzy coding framework to take into account the two steps. Prior incomplete knowledge indirectly deduced from administrative data in the design of the analysis is incorporated. As the database lacks labelled information, a pseudo labelled dataset to allow for an accurate evaluation of our method according to medical practices is retrieved and enhanced.

E1877: **Exploratory data analysis of sustainable development goals**
Presenter: **Vladimir Potashnikov**, RANEPA, Russia
Co-authors: Andrey Zubarev, Oleg Lugovoy

In September 2015, United Nations agreed on the agenda: "Transforming our World: The 2030 Agenda for Sustainable Development", which identified 17 sustainable development goals. At the same time, the agenda defines only indicators of sustainable development, but not aggregated indicators for each goal. The aim is to identify aggregated indicators of sustainable development goals (SDG) achievement that are suitable for clear economic interpretation, and suitable for analysis under which conditions that can be achieved. Another important goal is to identify groups of interrelated indicators of SDG, the objectives of which are more appropriate to consider together without other. This analysis was provided by hierarchical clustering methods. Also group of similarly countries in terms of achieved SDG was found based on these results. The possibility of achievement different couples of SDG indicators together was tested in this paper, based on historical data.

E0336: **Extreme value theory for open set classification: The GPD and GEV classifiers**
Presenter: **Edoardo Vignotto**, University of Geneva, Switzerland
Co-authors: Sebastian Engelke

Classification tasks usually assume that all possible classes are present during the training phase. This is restrictive if the algorithm is used over a long time and possibly encounters samples from unknown classes. It is therefore fundamental to develop algorithms able to distinguish between known and unknown new data. In the last few years, extreme value theory has become an important tool in multivariate statistics and machine learning. The recently introduced extreme value machine, a classifier motivated by extreme value theory, addresses this problem and achieves competitive performance in specific cases. However, this algorithm can fail when the geometries of known and unknown classes differ, even if the recognition task is fairly simple. To overcome these limitations, two new algorithms for open set classification relying on approximations from extreme value theory that are more robust in such cases are proposed. They exploit the intuition that test points that are extremely far from the training classes are more likely to be unknown objects. Asymptotic results motivated by univariate extreme value theory that make this intuition precise are proposed. The effectiveness of the new classifiers is shown in simulations and on real data sets.

E1702: **Bayesian sparse convex clustering via NEG distribution**
Presenter: **Kaito Shimamura**, NTT Advanced Technology Corporation / The University of Electro-Communications, Japan
Co-authors: Shuichi Kawano

Sparse convex clustering is convex clustering, which is a convex relaxation of classical clustering methods, with variable selection. Sparse regularization plays a key role of selecting relevant variables in sparse convex clustering. In sparse convex clustering, we need to set values of weights in the regularization term. By setting the values properly, it is known that the accuracy of clustering and variable selection improves. However, it is pointed out that the values highly depend on observed data. This causes a degradation of estimation accuracy in sparse convex clustering when the sample size is small. To overcome the problem, we first introduce a Bayesian formulation of sparse convex clustering. We then propose a Bayesian sparse convex clustering based on a normal-exponential-gamma (NEG) prior distribution. We conduct numerical studies to examine the effectiveness of the Bayesian model.

CI024 Room Beveridge Hall ECONOMETRICS OF VOLATILITY Chair: Mathieu Rosenbaum

C0376: **Volatility regressions with fat tails**
Presenter: **Nour Meddahi**, Toulouse School of Economics, France
Co-authors: Jihyun Kim

Nowadays, a common practice to forecast integrated variance is to do simple OLS autoregressions of the observed realized variance data. However, non-parametric estimates of the tail index of this realized variance process reveal that its second moment is possibly unbounded. In this case, the behavior of the OLS estimators and the corresponding statistics are unclear. We prove that when the second moment of the spot variance is unbounded, the slope of the spot variance's autoregression converges to a random variable when the sample size diverges. Likewise, the same result holds when one consider either integrated variance's autoregression or the realized variance one. We then consider a class of variance models based on diffusion processes having an affine form of drift, where the class includes GARCH and CEV processes, and we prove that IV estimations with adequate instruments provide consistent estimators of the drift parameters as long as the variance process has a finite first moment regardless of the existence of finite second moment. In particular, for the GARCH diffusion model with fat tails, an IV estimation where the instrument equals the sign of the (demeaned) lagged value of the variable of interest provides consistent estimators. Simulation results corroborate the theoretical findings.

C0163: **Estimation of the quadratic variation and its eigenvalues for multivariate jump processes**
Presenter: **Mark Podolskij**, Aarhus University, Denmark

New asymptotic theory for the estimation of quadratic variation in the setting of multivariate jump processes is presented. In financial applications quadratic variation plays a key role in the assessment of risk. In the past decade there have been numerous studies on statistical inference for quadratic variation. The mathematical theory depends very much on the particular modelling framework: when the jump part is absent the statistical theory is well understood since the 90's; if the Brownian and the jump parts are present the weak limit results, which have been investigated previously, are quite non-standard. We will consider the setting when the underlying semimartingale is a multivariate pure jump process. We will investigate the weak limit theory for the matrix-valued quadratic variation, present the corresponding results for its random eigenvalues and discuss how the theory can be applied in practice.

C0261: **From quadratic Hawkes processes to super-Heston rough volatility models with Zumbach effect**
Presenter: **Mathieu Rosenbaum**, Ecole Polytechnique, France

Using microscopic price models based on Hawkes processes, it has been shown that under some no-arbitrage condition, the high degree of endogeneity of markets together with the phenomenon of metaorders splitting generate rough Heston-type volatility at the macroscopic scale. One additional important feature of financial dynamics, at the heart of several influential works in econophysics, is the so-called feedback or Zumbach effect. This essentially means that past trends in returns convey significant information on future volatility. A natural way to reproduce this property in microstructure modeling is to use quadratic versions of Hawkes processes. We show that after suitable rescaling, the long term limits of these processes are refined versions of rough Heston models where the volatility coefficient is enhanced compared to the square root characterizing Heston-type dynamics. Furthermore the Zumbach effect remains explicit in these limiting rough volatility models.

CO741 Room Bloomsbury INFLATION EXPECTATIONS AND INFLATION DYNAMICS Chair: Francesco Grigoli

C0225: **The impact of guidance, short-term dynamics and individual characteristics on firms long-term inflation expectations**
Presenter: **Christian Raggi**, Swiss National Bank, Switzerland

Long-term inflation expectations are an essential element in the transmission of a central banks monetary policy. However, it is not fully understood how these long-term inflation expectations behave, as well as whether and how they can be managed. We shed light on possible drivers of long-term inflation expectations of firms using information from the Swiss National Bank (SNB) regional network survey. We extended the standard survey with questions to test whether the long-term inflation expectations of firms can be actively influenced by providing information regarding long-term average inflation, the central banks objective and past performance. We find that this type of information, which we call guidance: a) can influence the long-term inflation expectations of firms to a certain extent; and b) surprisingly, it does not have an impact on the uncertainty surrounding the expectations. However, c) uncertainty itself is positively correlated with the level of inflation expectations, and respondents who are more uncertain place greater weight on the information that they receive. Furthermore, d) short-term inflation expectations; e) the individual characteristics of the firms related to prices; and f) a large unexpected shock, in our case, a large shock to the exchange rate, can help to explain the behaviour of long-term inflation expectations.

C0226: **IQ, expectations, and choice**
Presenter: **Michael Weber**, Chicago Booth, United States

Administrative and survey-based micro data are used to study the relationship between cognitive abilities (IQ), the formation of economic expectations, and the choices of a representative male population. Men above the median IQ (high-IQ men) display 50% lower forecast errors for inflation than other men. The inflation expectations and perceptions of high-IQ men, but not others, are positively correlated over time. High-IQ men are also less likely to round and to forecast implausible values. In terms of choice, only high-IQ men increase their propensity to consume when expecting higher inflation as the consumer Euler equation prescribes. High-IQ men are also forward-looking - they are more likely to save for retirement conditional on saving. Education levels, income, socioeconomic status, and employment status, although important, do not explain the variation in expectations and choice by IQ. Our results have implications for heterogeneous-beliefs models of household consumption, saving, and investment.

C0243: **The role of expectations in changed inflation dynamics**
Presenter: **Damjan Pfajfar**, Board of Governors of the Federal Reserve System, United States
Co-authors: John Roberts

The Phillips curve has been much flatter in the past twenty years than in the preceding decades. We consider two hypotheses. One is that prices at the microeconomic level are stickier than they used to be—in the context of the canonical Calvo model, firms are adjusting prices less often. The other is that the expectations of firms and households about future inflation are now less well informed by macroeconomic conditions; because expectations are important in the setting of current-period prices, inflation is therefore less sensitive to macroeconomic conditions. To distinguish between our two hypotheses, we bring to bear information on inflation expectations from surveys, which allow us to distinguish changes in the sensitivity of inflation to economic conditions conditioning on expectations from changes in the sensitivity of expectations themselves to economic conditions. We find that, with some measures, expectations are less tied to economic conditions than in the past, and thus that this reduced attentiveness can account for a significant portion of the reduction in the sensitivity of inflation to economic conditions in recent decades.

C0343: **Monetary policy surprises and inflation expectation dispersion**
Presenter: **Francesco Grigoli**, International Monetary Fund, United States

The impact of monetary policy surprises on inflation expectation dispersion is analyzed. Relying on daily data of policy rate and inflation expectations at the analyst level for the United States and the United Kingdom, we calculate monetary policy surprises as unexpected changes in the policy rate two and a half days before the central bank meetings. This identification strategy isolates the effect of the information that is revealed by the central bank action (or inaction). The dispersion of inflation expectations is calculated, for different horizons, over the two and a half days following those meetings. We find that the information effect of the central bank decisions generates dispersion of inflation expectations at short horizons and has no effect or even a negative one at longer horizons. These results point to the need to improve communication of the monetary policy decisions so that dispersion is contained even at shorter horizons, when changes in policy rates do not have concrete effects.

CO528 Room G11 RECENT ADVANCES IN BAYESIAN MULTIVARIATE MODELLING AND ESTIMATION Chair: Michael Smith

C0201: **Bayesian inference for regression copulas**
Presenter: **Michael Smith**, University of Melbourne, Australia
Co-authors: Nadja Klein

A new semi-parametric distributional regression smoother for continuous data is proposed which is based on a copula decomposition of the joint distribution of the vector of response values. The copula is high-dimensional and constructed by inversion of a pseudo regression, where the conditional mean and variance are non-parametric functions of the covariates modeled using Bayesian splines. By integrating out the spline coefficients, we derive an implicit copula that captures dependence as a smooth non-parametric function of the covariates, which we call a regression copula. We derive some of its properties, and show that the entire distribution including the mean and variance of the response from the copula model are also smooth nonparametric functions of the covariates. Even though the implicit copula cannot be expressed in closed form, we estimate it efficiently using both Hamiltonian Monte Carlo and variational Bayes methods. We illustrate the efficacy of these estimators and copula model for implicit copulas up to dimension 40,981.

C0205: **Bayesian variable selection for non-Gaussian responses: A marginally calibrated copula approach**
Presenter: **Nadja Klein**, Humboldt University Berlin, Germany
Co-authors: Michael Smith

A new highly flexible and tractable Bayesian approach is proposed to undertake variable selection in non-Gaussian regression models. It uses a copula decomposition for the vector of observations on the dependent variable. This allows the marginal distribution of the dependent variable to be calibrated accurately using a nonparametric or other estimator. The family of copulas employed are 'implicit copulas' that are constructed from existing hierarchical Bayesian models used for variable selection, and we establish some of their properties. Even though the copulas are high-dimensional, they can be estimated efficiently and quickly using Monte Carlo methods. A simulation study shows that when the responses are

non-Gaussian, the approach selects variables more accurately than contemporary benchmarks. A marketing example illustrates that accounting for even mild deviations from normality can lead to a substantial improvement. To illustrate the full potential of the approach, we extend it to spatial variable selection for fMRI data. It allows for voxel-specific marginal calibration of the magnetic resonance signal at over 6000 voxels, leading to a considerable increase in the quality of the activation maps.

C0218: **High-dimensional copula variational approximation through transformation**
Presenter: **Ruben Loaiza-Maya**, Monash University, Australia
Co-authors: Michael Smith, David Nott

Variational methods are attractive for computing Bayesian inference for highly parametrized models and large datasets where exact inference is impractical. They approximate a target distribution - either the posterior or an augmented posterior - using a simpler distribution that is selected to balance accuracy with computational feasibility. We approximate an element-wise parametric transformation of the target distribution as multivariate Gaussian or skew-normal. Approximations of this kind are implicit copula models for the original parameters, with a Gaussian or skew-normal copula function and flexible parametric margins. A key observation is that their adoption can improve the accuracy of variational inference in high dimensions at limited or no additional computational cost. We consider the Yeo-Johnson and G&H transformations, along with sparse factor structures for the scale matrix of the Gaussian or skew-normal. We also show how to implement efficient reparametrization gradient methods for these copula-based approximations. The efficacy of the approach is illustrated by computing posterior inference for three different models using six real datasets. In each case, we show that our proposed copula model distributions are more accurate variational approximations than Gaussian or skew-normal distributions, but at only a minor or no increase in computational cost.

C0222: **Bayesian estimation and testing for constrained multivariate functions**
Presenter: **Tom Shively**, University of Texas at Austin, United States

The aim is to estimate multivariate functions nonparametrically with shape constraints such as monotonicity, convexity and quasi-convexity imposed on the function estimates. We also develop tests for whether it is appropriate to impose specific shape constraints in some or all directions. Our method uses a regression spline representation of the multivariate function and projects the unconstrained spline function into the appropriate constrained function space. Quadratic programming is used to solve for the constrained regression spline coefficients. Simulation experiments show the small sample properties of both the estimation and testing methodology.

CO414 Room G4 MODELING REGIME CHANGE I Chair: Willi Semmler

C1525: **Credit risk and delayed monetary policy effectiveness: A finite horizon multi-phase model**
Presenter: **Willi Semmler**, New School for Social Research, United States

Given the long period of expansionary monetary policies following the great recession 2008-9, many observers claim that those policies exerted their effects on the real economy through the asset market and repricing of credit risk. To study this channel we add nonlinear dynamics for credit flows and credit spreads in a regime switching inflation targeting model and explore the stabilizing - destabilizing effects of the dynamics of credit flows and risk premia. We study the effectiveness of conventional and unconventional monetary policies on credit flows and credit risk under simultaneous and delayed policy impacts in a multi-phase dynamic model. We use estimated parameters, based on data for the Euro area, and solve a nonlinear controlled dynamic system through AMPL for a finite horizon model. We find that with longer delays policies might not be able to effectively stabilize the macro imbalances in particular if a regime switch has occurred. Though in our context the agents maybe forward looking over a finite horizon, there are also echo effects from the past that come into play with a delay affecting real and financial variables.

C1495: **State-dependent effects of monetary policy: The central bank information channel**
Presenter: **Paul Hubert**, Sciences Po - OFCE, France

When the central bank and private agents do not share the same information, private agents may not be able to appreciate whether monetary policy responds to changes in the macroeconomic outlook or to changes in policy preferences. In this context, we investigate whether the publication of the central bank macroeconomic information set modifies private agents interpretation of policy decisions. We find that the sign and magnitude of the effects of monetary policy depend on the publication of policymakers macroeconomic views. Contractionary monetary policy has negative effects on inflation expectations and stock prices only if associated with inflationary news.

C1459: **The new Eurozone risk morphology**
Presenter: **Marcello Minenna**, CONSOB, Italy

Ten years into the global financial crisis, the euro area is struggling to get back on a path of stability and growth. Leaving aside international factors, the underlying reasons come from within, ranging from the EMU architectural incompleteness to the reluctance to address some key issues, starting from the ECB mandate and constraints. These reasons develop along two main risk backbones that define the Eurozone risk morphology: large and persistent competitive gaps, which contrast center and periphery, and systematic risk segregation, which hinders effective progress towards a fiscal union. These two risk backbones are explored and measured through economic and financial indicators that are closely related to each other. The critical values of these indicators highlight a matter of unsustainability of the EMU membership, as hinted by the rising Euro-skeptic debate. This has resulted in a confrontation attitude of most distressed countries with the European institutions, which in turn has been translated into higher sovereign risk premia as in the recent Italian experience. The recipe for these problems cannot be limited to strengthened budgetary surveillance and stability discipline of the financial sector: it must open to risk sharing in order to definitively defuse centrifugal forces, remove financial and commercial imbalances, and pave the way for a fiscal union with a federal budget, an unified debt market and a single finance minister.

C1549: **Economic dynamics in the presence of catastrophic risk: The danger of poverty traps**
Presenter: **Raimund Kovacevic**, Vienna University of Technology, Austria
Co-authors: Willi Semmler

The focus is on economic dynamics described by an optimal control problem, which shows three equilibrium points: An unstable equilibrium lies between an upper and a lower stable equilibrium. In the presence of rare but substantial random shocks between phases of deterministic growth, we analyze the risk of ending up below the unstable equilibrium, which can be considered as poverty trap. We also aim at decision making within this setup, particularly considering the role of insurance against catastrophic events.

CO847 Room G5 BUSINESS CYCLE ANALYSIS Chair: Michael Owyang

C0410: **Tax progressivity, economic booms, and trickle up economics**
Presenter: **Christopher Otrok**, University of Missouri and FRB St Louis, United States
Co-authors: Michael Owyang, Laura Jackson Young

An increase in tax progressivity sets off an economic boom. Those at the bottom of the income distribution (who are constrained hand to mouth consumes) set off a consumption boom that expands the overall economy. Those at the top of the income distribution disproportionality benefit from expansions, and their income gains from the boom more than offset the increases in tax from higher marginal rates. The empirical results show that aggregate income and consumption rise after an increase in progressivity. At the same time the income Gini rises (as do other inequality measures comparing percentiles of income). We interpret these results as evidence in favor of tickle up, not trickle down, economics. Such a policy also has no impact on deficits and raises the tax revenue to GDP ratio in the longer run, which we interpret as due to the economic expansion.

A methodological novelty is a new measure of income tax progressivity. We propose a method to decompose changes in the tax structure into a component measuring the level of taxes and a component orthogonal to the level that measures progressivity. While the focus is on the progessivity results, we find that the level shock is similar to the standard tax shocks that are found in the empirical literature in that a rise in the level is contractionary.

C0628: **Networking the yield curve**
Presenter: **Julia Schaumburg**, VU University Amsterdam, Netherlands
Co-authors: Tatjana Dahlhaus, Tatevik Sekhposyan

The term structure of interest rates and its dynamics across the business cycle are studied. Relative to the empirical literature on modeling the yield curve, it addresses the contemporaneous cross-correlations between the yields of different maturities, above and beyond what could be modeled by assuming a common factor structure across the yields. The contributions are twofold. First, we propose a spacial time series modeling framework for the yield curve and investigate its performance in a context of simulations. We further apply the model to the Treasury yields in the US and link the dynamics of the contemporaneous correlations and their evolution to the business cycle and the conduct of monetary policy.

C0382: **A score-driven model for GDP-at-risk**
Presenter: **Andrea De Polis**, University of Warwick, United Kingdom
Co-authors: Davide Delle Monache, Ivan Petrella

A fully parametric model is proposed to characterize the predictive density of GDP growth. In our trend-cycle model, the disturbances follow an Epsilon Skew-t distribution with time-varying moments (location, scale and shape), whose dynamics is driven by the score of the predictive likelihood and possibly by additional exogenous information. When we include financial condition indices as additional drivers in the updating processes, we observe significant improvements in the out of sample predictive ability for different horizons reflecting the model's ability to pick up in a timely manner changes in the shape of the forecast density. We also recover most of the stylized facts about GDP growth documented in literature. Particular attention is devoted to GDP vulnerability as proxied by the asymmetry of the predictive distribution. We find that financial tightening are robust drivers of left skewness of the predictive distribution, ultimately sharpening economic growth predictions at the onset of recessions.

C1489: **The macroeconomic effects of bank capital regulation**
Presenter: **Benedikt Kolb**, Deutsche Bundesbank, Germany
Co-authors: Esteban Prieto, Sandra Eickmeier

Bank capital regulations aim at reducing risk-taking and increasing the resilience of the financial sector. A key concern, however, is whether higher capital requirements impair banks' ability to lend, with potentially long-lasting negative effects for the economy. We propose a narrative identification strategy to examine the macroeconomic effects of higher bank capital requirements. We exploit the staggered implementation of these policies to account for anticipation effects of changes in capital regulation. We find that higher capital requirements lead to a sizable reduction in bank assets and lending, with substantial negative spillovers to the real economy. These effects are, however, only short-lived and temporary. We do not find evidence for long-run negative effects of higher capital requirements on bank lending and economic activity.

CO200 Room Gordon TIME SERIES ECONOMETRICS Chair: Antonio Montanes

C0438: **On what drives regional business cycles in Europe**
Presenter: **Ana Gomez-Loscos**, Bank of Spain, Spain
Co-authors: Lola Gadea, Eduardo Bandres

The aim is twofold. First, to identify regional business cycles in Europe, obtaining datings of business cycles that allow us to identify possible groups of regions and to assess how interdependence in European regions has changed over time, paying special attention both to the euro cash changeover and the global financial crisis. Second, to determine the driving forces behind each group of regions, that is, to identify the economic, cultural and social factors that may help explain regional business cycles. We identify five different groups of European regions. Moreover, we find an increase over time in regional business cycle synchronization. We detect that the variables that play the most important role to identify groups of regions are those related to well-being and geography and culture.

C0691: **Time-varying behavior of the equilibrium velocity of money in the Euro area**
Presenter: **Juan Sapena**, Catholic University of Valencia, Spain
Co-authors: Mariam Camarero, Cecilio Tamarit

The historical importance of money growth on inflation changes has been well established in the literature. This is the reason why the study of the stability of the money demand function is of such great importance. Nevertheless, for this function to be stable, the velocity of circulation of money should not change or, at least, its deviations from its long-run value should not be permanent. However, recent developments in inflation and in M3 velocity in the Euro area have raised serious doubts about the reliability of M3 growth as a pillar of the ECBs monetary policy strategy. We develop a very flexible and comprehensive state-space framework for modeling the velocity of circulation. Our specification allows for the estimation of different autoregressive alternatives, and include control instruments, whose coefficients can be set-up either common or idiosyncratic. This is particularly useful to detect asymmetries among individuals (countries) to common shocks.

C0951: **Long-term climate forecasts**
Presenter: **Lola Gadea**, University of Zaragoza, Spain
Co-authors: Jesus Gonzalo Munoz

Climate is a long-run phenomenon, and then the interesting forecast should convey certain long-term flavour (5, 10, 25, 50, 100 years ahead). The existence of a trend locally (time series data from Central England) as well as globally (cross-section stations from NH) in all the temperature distributional characteristics (not only the mean) has been previously detected. This is a definition of local or global warming. What it was not provided is the type of trend those distributional characteristics contain. We propose different trend models to capture the evolution of the temperature distribution and, by running a forecast competition, we choose the trend model that forecasts best. We also provide temperature long-run forecast from the best model, as well as from a combination of forecasts. One of the problems of long-run forecast is the length of the confidence intervals. We propose to construct the forecast confidence intervals by forecasting the different quantiles of the temperature distribution. These intervals are smaller than the ones constructed by standard forecast methods. These forecasts provide another picture of how serious is the local-global warming.

C1093: **Testing for trends in the presence of seasonal component**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain

A method is proposed for testing the presence of a trend in series that contain seasonal component. This method is an extension of a previous one. We derive the asymptotic behavior and analyze the finite sample properties by way of some Monte Carlo simulations. We illustrate its use by studying the presence of a trend in some temperature series.

CO220 Room Montague TOPICS IN FINANCIAL ECONOMETRICS I Chair: Leopold Soegner

C0801: **Varying correlation parametrizations in an HMM setting for filter-based portfolio strategies**
Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany
Co-authors: Stefanie Grimm, Peter Ruckdeschel, Joern Sass, Tilman Sayer

Portfolio optimization is considered in a regime-switching market. The assets of the portfolio are modeled through a hidden Markov model (HMM) in discrete time, where drift and volatility of the single assets are allowed to switch between different states. We consider different parametrizations of the involved asset covariances namely state-wise uncorrelated assets (though linked through the common Markov chain), assets correlated in a state-independent way, and assets where the correlation varies from state to state. As a benchmark we also consider a model without regime switches. We utilize a filter-based EM-algorithm to obtain optimal parameter estimates within this multivariate HMM and present parameter estimators in all three HMM settings. We discuss the impact of these different models on the performance of several portfolio strategies. Our findings show that for simulated returns our strategies in many settings outperform naïve investment strategies, like the equal weights strategy. Information criteria can be used to detect the best model for estimation as well as for portfolio optimization. A study using real data confirms these findings.

C0883: **Performance of equal weight strategies using fewer assets**
Presenter: **Anna-Katharina Thoes**, TU Kaiserslautern, Germany
Co-authors: Joern Sass

Diversification is one of the main pillars of investment strategies. The prominent 1/N-portfolio, which puts equal weight on each asset, is apart from its simplicity a method which is hard to outperform in realistic settings. But depending on the number of considered assets this method can lead to very large portfolios. We investigate how the number of assets can be reduced and which advantages and disadvantages arise with such a reduction. Therefore we investigate different naive portfolios from selecting the best Sharpe ratio assets to exploiting knowledge about correlation structures using clustering methods. The clustering techniques separate the possible assets into non-overlapping clusters and the assets within a cluster are ordered by their Sharpe ratio. Then the best asset of each portfolio is chosen to be a member of the new portfolio with equal weights, the cluster portfolio. We show that this portfolio inherits the advantages of the 1/N-portfolio and can even outperform it empirically. We finally derive corresponding results in comonotonic model settings for the clusters, and show how they can explain our observations on real data.

C1402: **Analyzing and testing the forward bias puzzle**
Presenter: **Julia Reynolds**, Universita della Svizzera italiana, Switzerland
Co-authors: Leopold Soegner, Martin Wagner

New econometric tools are applied to explore the so-called "forward bias puzzle", or the finding that the forward unbiasedness hypothesis, which states that forward exchange rates should predict expected future spot rates, fails to hold in real-world datasets. Modelling the forward unbiasedness hypothesis as a cointegrating regression allows us to apply previous monitoring tools, which monitor the stability of a cointegrating regression over time. The results show that deviations from the forward unbiasedness hypothesis are much more likely at longer maturities, implying a potential liquidity effect in the forward bias puzzle.

C1083: **Size and power properties of autocorrelation and heteroskedasticity robust tests in spatial error models**
Presenter: **Christian Zwatz**, University of Vienna, Austria

A typical approach for testing linear hypotheses on the regression parameters in regression models with autocorrelated and/or heteroskedastic disturbances is to modify the conventional F-test statistic by using a heteroskedasticity and autocorrelation consistent (HAC) estimator for the covariance matrix. These are nonparametric estimators designed to take the heteroskedasticity and autocorrelation in the data into account. We consider heteroskedasticity and autocorrelation robust testing in spatial error models, i.e. models where the disturbances follow a spatial autoregressive or spatial moving average process. It is well known that tests based on HAC estimators in the case of time series regression models suffer from substantial size and power problems. Based on a general theory about size and power properties of tests in regression models with autocorrelated and/or heteroskedastic disturbances, we show that similar problems also occur in the spatial error model. In particular, we give conditions under which the size of the resulting test is in fact one. We also give conditions under which the size of the test can be controlled by an appropriate choice of critical value.

CO224 Room Woburn TOPICS IN DYNAMIC MACROECONOMICS AND MACROECONOMETRICS Chair: Marco Maria Sorge

C0299: **Macroeconomic outcomes in disaster-prone countries**
Presenter: **Alessandro Cantelmo**, Bank of Italy, Italy
Co-authors: Giovanni Melina, Chris Papageorgiou

Using a dynamic stochastic general equilibrium model, we study the channels through which natural disaster shocks affect macroeconomic outcomes and welfare in disaster-prone countries. We solve the model using Taylor projection, a solution method that is shown to deal effectively with high-impact weather shocks calibrated in accordance to empirical evidence. We find large and persistent effects of weather shocks that significantly impact the income convergence path of disaster-prone countries. Relative to non-disaster-prone countries, on average, these shocks cause a welfare loss equivalent to a permanent fall in consumption of 1.6 percent. Welfare gains to countries that self-finance investments in resilient public infrastructure are found to be negligible, and international aid has to be sizable to achieve significant welfare gains. In addition, it is more cost-effective for donors to contribute to the financing of resilience before the realization of disasters, rather than disbursing aid after their realization.

C0311: **One size to fit all in the Euro area: Some counterfactual evidence**
Presenter: **Matteo Fragetta**, University of Salerno, Italy
Co-authors: Sergio Destefanis, Emanuel Gasteiger

The aim is to examine whether Euro Area countries would have faced a more favorable inflation output variability tradeoff without the Euro. We provide evidence that this claim is true for the periods of the Great Recession and the European Sovereign Debt Crisis. For the Euro Area as a whole, the deterioration of the tradeoff becomes insignificant with Draghi's whatever it takes announcement onwards. However, a more detailed analysis shows that the detrimental effect of the Euro is more severe and long-lasting for peripheral countries, pointing to structural differences among Euro Area countries as a key element of the detrimental effect of the Euro. We base our results on a novel empirical strategy that, consistently with monetary theory, models the joint determination of the variability of inflation and output conditional on structural supply shocks. Moreover, our findings are robust to potential endogeneity concerns related to adoption of the Euro.

C0584: **Synthesizing structural evidence on the monetary and fiscal stance in the US: A Bayesian approach**
Presenter: **Andreas Tryphonides**, University of Cyprus, Cyprus

The identification of policy regimes from macroeconomic data almost always relies on particular identifying assumptions. We instead synthesize evidence based on alternative assumptions or methodologies. We treat identified shocks using heterogeneous restrictions and other methodologies i.e. narrative schemes as imperfect observations on the true latent shocks. We illustrate how posterior estimates of the latent state update on new information where the latter depends on how credible the identifying restrictions are. Based on this methodology, we identify monetary and fiscal policy regimes in the United States. A by-product of the methodology is an MCMC algorithm for conducting joint quasi-Bayesian inference on reduced form and structural parameters, useful for limited information structural equations with unobservables and possibly time varying parameters.

C0759: **Arbitrary initial conditions and the dimension of indeterminacy in linear rational expectations models**
Presenter: **Marco Maria Sorge**, University of Salerno, Italy

Indeterminate equilibrium rational expectations (RE) models are ubiquitous in both theoretical and applied work in dynamic macroeconomics. The issue of characterizing the exact dimension of indeterminacy - i.e. of deriving the full set of causal and stable solutions to linear RE models - has only recently been addressed in the context of general, multivariate settings. Existing results are complemented by identifying bounds on the observable dimension of indeterminacy of linear RE models in the presence of arbitrary initial conditions. In particular, it is established that, provided the underlying RE model admits a non-unique (causal, stable) solution, then (i) the exact dimension of indeterminacy is always lower than (or at most equal to) the degree of indeterminacy as previously identified, and (ii) the maximal dimension of indeterminacy cannot exceed the one associated with the model's counterpart featuring initial conditions which are set to lie onto the model's stable saddle path. Implications for the estimation of indeterminate equilibrium RE models are discussed.

CO202 Room Chancellor's Hall HIGH-FREQUENCY ECONOMETRICS Chair: Bezirgen Veliyev

C1153: **A nonparametric test for commonality in intraday high-frequency data**
Presenter: **Kim Christensen**, Aarhus University, Denmark

A nonparametric test is developed to detect the presence of diurnal variation in the correlation coefficient between asset price processes in intraday high-frequency data. A simulation study shows the test has good size and power properties. An empirical illustration shows the advantage of taking intraday correlation into account.

C1020: **Information flows and volatility: Level or persistence shifts**
Presenter: **Daniel Borup**, Aarhus University, Denmark
Co-authors: Kristoffer Pons Bertelsen, Johan Jakobsen

The relationship between information flows and the dynamic properties of financial market volatility is examined. In order to do so, we formulate a smooth-transition realized GARCH framework where the level and persistence parameters are allowed to vary, possibly jointly, over time as a function of news arrival. We entertain the possibility of a non-linear relationship. The model dynamically controls for effects of realized volatility on the conditional variance process which otherwise may bias conclusions. Our findings on a broad panel of international stock market indices indicate a strong relationship between both the level and persistence with the strength of news arrival. Negatively loaded information flows matter in particular. This relationship is non-linear, showing a clear bell-shaped structure. That is, weak or strong information arrival characterize relatively low level and persistence states of volatility, whereas medium information flows characterize high level and persistence states of volatility. Our findings may be explained by the degree of ambiguity in the signals of incoming information and market participants' disagreement of their interpretation.

C0600: **Forecasting the volatility of the yield curve**
Presenter: **Mads Markvart Kjaer**, Aarhus University, Denmark
Co-authors: Bezirgen Veliyev, Bent Jesper Christensen

The ability of various forecasting methods to forecast future volatilities of the yield curve is examined. We document that standard affine term structure models are not able to capture all relevant information from the yield curve about future volatility and we find evidence for unspanned stochastic volatility even when we take the information from the yield curve into account. Overall, we conclude that it is difficult to significantly out-perform a random walk and that time-series approaches provide lower root-mean-squared forecast errors. We further find that changing the starting point of the out-of-sample has a large impact on the conclusions and, hence, we find switching behavior of the predictive-ability of term structure models regarding future volatility.

C0638: **A regime-switching stochastic volatility model for forecasting electricity prices**
Presenter: **Peter Exterkate**, University of Sydney, Australia
Co-authors: Oskar Knapik

Three crucial challenges outstanding in the area of electricity price forecasting are addressed. Specifically, we show the importance of considering fundamental price drivers in modelling, develop new techniques for probabilistic (i.e. interval or density) forecasting of electricity prices, and introduce a universal Bayesian technique for model comparison. We propose a new regime-switching stochastic volatility model with three regimes, which may be interpreted as negative jump or "drop", normal price or "base", and positive jump or "spike", respectively. The transition matrix between these regimes is allowed to depend on explanatory variables in a novel way, using an underlying ordered probit model. Bayesian inference is employed in order to obtain predictive densities. The main focus is on short-term density forecasting in the Nord Pool intraday market. We show that the proposed model outperforms several benchmark models at this task, as measured by their predictive Bayes factors. In particular, the incorporation of stochastic volatility, regime switching, information from the day-ahead market, and exogenous information from weather reports into the model are all shown to improve its predictive performance, without falling prey to curse of dimensionality problems.

CC816 Room MAL 351 CONTRIBUTIONS IN FINANCIAL TIME SERIES Chair: Robert Jung

C0865: **Estimation of the parameters of symmetric stable ARMA and ARMA-GARCH time series process**
Presenter: **Aastha Madonna Sathe**, IIT MADRAS, India
Co-authors: Neelesh Shankar Upadhye

The modified Hannan-Rissanen Method is proposed which is useful in estimating the parameters of symmetric stable autoregressive moving average (ARMA) time series process. The proposed method is also effective in estimating the ARMA process with the symmetric stable generalized autoregressive conditional heteroskedasticity (GARCH) noise. The efficiency, accuracy and simplicity of our method is shown through Monte-Carlo simulation. Finally, as an application of the proposed method, we model and assess the financial heavy-tailed data to show its practicality.

C1716: **Public attention and financial markets: An analysis using google trends**
Presenter: **Shehroz Azmat**, University of Aberdeen, United Kingdom
Co-authors: Marc Gronwald, Xin Jin

The relationship between public attention and financial markets such as the markets for crude oil and the cryptocurrency Bitcoin is considered. Different google search terms are used in order to capture different aspects of public attention: while a term such as oil price captures general interest in this market, oil supply reflects interest in market fundamentals, and oil price bubble specific interest during extreme price episodes. Using Granger causality tests, it is found that these different measures of public attention are driven by market price movements in different ways; while e.g. the crude oil price Granger causes the search term oil market, no such relationship is found between oil prices and the search term oil bubble.

C1618: **Long-run predictability tests are even worse than you thought**
Presenter: **Tamas Kiss**, Orebro University, School of Business, Sweden
Co-authors: Erik Hjalmarsson

The interaction between the two problems of endogenous predictors and inference in long-horizon regressions is studied. The key finding is that long-horizon predictive regressions exacerbates the endogenous predictor bias. Specifically, while endogenous predictors are usually considered

problematic only if they are sufficiently persistent, we show that in long-horizon regressions, a version of the Stambaugh bias is present - and substantial - regardless of the (lack of) persistence in the predictor. We derive asymptotic results for a scaled version of the OLS t-statistic. With exogenous regressors, the scaling correctly controls for the overlap in the data, and the scaled t-statistic is very close to normally distributed also in finite samples. With endogenous regressors, the distribution of the scaled t-statistic differs substantially from standard normal. This holds regardless of the persistence in the predictor, and the (asymptotic) Stambaugh bias arising in the scaled t-statistic is thus independent of the persistence in the predictor, and completely induced by the formulation of the long-horizon regression.

C0192: **Intraday conditional value at risk: A periodic mixed-frequency GAS approach**
Presenter: **Bastian Gribisch**, University of Cologne, Germany
Co-authors: Tobias Eckernkemper

A copula-based periodic mixed frequency GAS framework is proposed in order to model and forecast the intraday Exposure Conditional Value at Risk (ECoVaR) for an intraday asset return and the corresponding market return. In particular we analyze GAS models which account for long-memory-type of dependencies, periodicities, asymmetric nonlinear dependence structures, fat-tailed conditional return distributions and intraday jump processes for asset returns. We apply our framework in order to analyze the in-sample and out-of-sample ECoVaR forecasting performance for a large data set of intraday asset returns of the S&P500 index.

CC820 Room MAL 352 CONTRIBUTIONS IN VALUE-AT-RISK Chair: Alexandra Dias

C1648: **The impact of ESG on stocks downside risk and risk adjusted return**
Presenter: **Andreas Stephan**, Jonkoping University, Sweden
Co-authors: Hans Loof, Maziar Sahamkhadam, Andreas Stephan

Investments considering corporate social responsibility continue to expand. Are companies pursuing a CSR agenda benefiting shareholders by reducing their financial downside risk? The aim is to investigates the relationship between a firm's environmental, social and corporate governance (ESG) scores and its downside risk on the stock market. We study this link using a panel of 887 stocks listed in five European countries over the period 2005-2017. Our empirical results show that higher ESG scores are associated with reduced downside risk of stock returns. Based on the Fama-French three factor model, we found no systematic relationship between ESG and the level of risk-adjusted return.

C1614: **Asymmetric-loss-based evaluation of daily value-at-risk models**
Presenter: **Anna Titova**, Institute for Statistics and Econometrics, Germany

A comprehensive comparison of models for forecasting daily value-at-risk is presented. While most of the similar studies perform such analyses using only a few financial time series, the main goal is to rank forecasting performances of a multitude of models on a substantially larger dataset. The models are ranked according to statistical as well as regulatory criteria with guidelines suggested by the Basel accords. Modeling value-at-risk as a conditional quantile via heterogeneous quantile autoregression has shown the best overall results. Additionally, including external predictors containing market characteristics improves a models performance. The validity of the conclusions for expected shortfall forecasts is examined.

C1836: **Earning at risk for electricity generators**
Presenter: **Lin Han**, Macquarie University, Australia
Co-authors: Nino Kordzakhia, Stefan Trueck, Karol Binkowski

A numerical method for evaluation of the Earnings at Risk (EaR), a cashflow based volumetric risk measure driven by demand for electricity, is developed. A class of Normal Inverse Gaussian (NIG) distributions is employed to model perturbations of the load and electricity spot price processes at the intraday tick times. The algorithm for evaluating the distribution of a product of two NIG distributed random variables using the technique implied by solving the problem in the conditional Gaussian case. The conditional non-central chi-squared distribution plays a central role in the evaluation of the product of Gaussian variables. The bivariate NIG distribution allows addressing heavy tailedness of the electricity spot price and load variables at the intraday tick times. The performance of the suggested method is tested via Monte Carlo simulation using historical data supplied by the Australian Energy Market Operator.

C1804: **Scoring function-based model risk of risk models**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Ning Zhang, Radu Tunaru

Studying the accuracy of risk measures, we propose a scoring function-based model risk estimation methodology, which opens the possibility of measuring joint model risk of the pair of risk measures (VaR, ES) at significance level. We carry out a simulation study to illustrate and analyze our proposed model risk measure across various (VaR, ES) risk models, and also study the properties of the proposed methodology as a measure of (model) risk. An empirical analysis illustrates its application for different asset classes.

CC824 Room MAL 353 CONTRIBUTIONS IN COMPUTATIONAL ECONOMETRICS Chair: Peter Winker

C1882: **Correcting the bias of economic aggregates that is caused by classification errors**
Presenter: **Quinten Meertens**, University of Amsterdam, Netherlands
Co-authors: Cees Diks, Jaap van den Herik, Frank Takes

In economic statistics, estimated aggregates are often based on underlying classifications. If the class labels are predicted by a classification algorithm, the data may contain classification errors. It occurs, for example, when social media data are used to estimate the number of people that will vote for a political candidate. The focus in on the effect of classification errors on the accuracy of estimated aggregates (such as counts). The first finding was that even highly accurate classification algorithms might result in relatively strongly biased aggregates. Then, we developed novel methods to correct that bias, making more effective use of accuracy data such as estimated precision and recall (or estimated type I and type II error rates). The new methods are shown to have serious implications for a wide range of applications in economics and machine learning, including e-commerce estimates, land use statistics, epidemiology and elections predictions. Currently, we look into the potential of ranking over classification and algorithm-specific bias corrections. The aim is to develop bias corrections at the micro-level that lead to a minimized mean-squared error on the aggregate level.

C1924: **Model calibration and validation via confidence sets**
Presenter: **Raffaello Seri**, University of Insubria, Italy
Co-authors: Mario Martinoli, Samuele Centorrino, Davide Secchi

An earlier calibration and validation method for simulation models is extended. The previous method was based on the concept of Model Confidence Set, as is the present one. Given a distance between time series, a benchmark dataset, and a finite set of simulation models $\mathcal{M}$, the method allowed the researcher to build a confidence set, obtained as a subset of $\mathcal{M}$, containing, with prescribed probability, the model (or models) minimizing the distance with respect to the data. The main drawback of the method was that, in accordance with most approaches to calibration, it neglected the variability of the data and focused on the simulations. We investigate the effects of the variability of the data on the procedure and, when necessary, we propose some modifications.

C1696: **Recursive estimation for high-dimensional state-space model**
Presenter: **Shaowen Liu**, University of Padova, Italy

Co-authors: Massimiliano Caporin, Sandra Paterlini

The aim is to investigate and compare the performance of several types of recursive inference algorithms, including stochastic ensemble Kalman filter (SEKF), ensemble transformation Kalman filter (ETKF) and particle filter (PF), within high-dimensional state-space models, such as the linear Gaussian state-space model and the mixed-Gaussian state-space model. The latter case is especially useful when the true state is a 2D spatial field. In our simulations, we design several field patterns as the hidden state, and examine how the algorithms recover the latent pattern. As for parameter estimation, MCMC is applied. Based on the idea of particle marginal Metroplis-Hastings (PMMH), we propose to use ETKF, instead of particle filter, to generate proposal distribution in each MCMC step. We show that ETKF might be a better choice in case of high-dimensional state-space model.

C0187: **Computation of interval forecast for ARIMA models accounting for the uncertainty of parameters' estimates**

Presenter: **Nikita Moiseev**, Plekhanov Russian University of Economics, Russia

Co-authors: Nikolay Tikhomirov

A numeric method is introduced for calculating the confidence interval for ARIMA type models, taking into account the uncertainty of parameters' estimation, what is especially important with a relatively short data frame. The proposed method is based on estimating the model parameters through the Yule-Walker system of equations, which uses autocorrelation coefficients of various orders. A method is presented for estimating the variance-covariance matrix for the autocorrelation coefficients with the subsequent application of the numerical method for estimating the variance of the regression function. In addition to the theoretical calculations, simulations are carried out to test the developed method in comparison with the traditional method of obtaining the interval forecast.

Sunday 15.12.2019 10:50 - 12:55 Parallel Session H – CFE-CMStatistics

EO300 Room CLO B01 APPLIED FUNCTIONAL DATA ANALYSIS Chair: Marzia Cremona

E0258: **Revealing technical trading rules with the empirical similarity concept**
Presenter: **Yarema Okhrin**, Universitaet Augsburg, Germany

Economists frequently suggest formal mathematical models or theories that postulate a formal decision rule. These rules can be subsequently calibrated and validated using empirical data. When there is no suitable rule at hand, economic agents may rely on so-called case-based inference, which grounds on analyzing cases by drawing analogies between experienced situations and their outcomes and their analogy to the present problem. The empirical similarity concept puts the case-based decision theory into an econometric framework, by predicting the variable of interest by a sum of historical outcomes weighted by the distances between the current levels of covariates and their historical counterparts. We formalize the tools of technical analysis aimed to forecast asset prices and returns. In contrary to mostly heuristic definition of patterns in technical analysis, we rely on the B-spline regularization to define and classify the patterns. This technique can also be seen as a decomposition of asset prices and a parametric quantification of patterns typical in technical analysis. The empirical application shows that the obtained patters deviate from the patterns commonly assumed in practice. Thereafter, we apply empirical similarity to compare the historical patters with the current one and to build a forecast for the future price/return dynamics. We compare the approach to common benchmarks.

E0867: **Functional additive regression for ordinal responses: Modeling EEG- based brain arousal state dynamics**
Presenter: **Fabian Scheipl**, Ludwig-Maximilians-Universitaet Muenchen, Germany
Co-authors: Juliane Minkwitz

The aim is to present an extension of penalized likelihood-based and boosting-based generalized additive models for functional responses to ordinal functional responses, i.e., multiple time series of ordinal measurements, and to show how to embed this problem in the general framework of functional regression models. The models are applied to high-frequency ordinal time series of vigilance levels derived from resting state EEG recordings in order to quantify potential associations between depressive symptoms and the temporal dynamics of brain arousal regulation.

E0923: **A functional regression model for determining drug-response relationship in cancer genomics**
Presenter: **Juhyun Park**, Lancaster University, United Kingdom
Co-authors: Evanthia Koukouli, Frank Dondelinger

Cancer is a complex disease caused by abnormal cell growth in the organism. Despite its diversity in appearance, it is suspected that there may be some common biological processes that govern the underlying dynamics. Experiments using cell lines derived from patients offer valuable resources in this regard. We investigate the effectiveness of six anticancer compounds applied to different cancer cell lines under different dosage levels, adjusting for the genetic profiles of individual subjects under treatment. We assume that out of tens of thousands genes regulating proteins composition only a small proportion is actually associated with cancer cells survival in a dosage-dependent manner. Using the longitudinal and transcriptional data from the Genomics of Drug Sensitivity in Cancer (GDSC) project, we build a dose-varying regression model, a type of functional regression model with functional responses. The covariates include dose-invariant factors but their effects are assumed to vary smoothly over the dosage levels. Due to the dimensionality of the covariates, we combine a nonparametric screening with the selection of variables using a penalised regression. We evaluate the effectiveness of our method using simulation studies, focusing on the choice of the thresholds for screening and penalty parameters.

E1320: **Mixture of functional graphical models with an application to ADHD data**
Presenter: **Hyun Bin Kang**, Western Michigan University, United States
Co-authors: Qihai Liu, Lucas Price, Kevin Lee

Many scientific areas are faced with the challenge of extracting information from large complex data. As a part of such effort, a functional graphical model is developed to extract the conditional dependence structure among random functions. The mixture of functional graphical models is presented, where the random functions are from a mixture of Gaussian processes with different conditional dependence structures. An EM algorithm for finding the mixture group of each function and the conditional dependence structure will be proposed, and a BIC-based criteria for selecting an optimal number of groups will be presented. The motivation comes from an ADHD dataset, which will be discussed.

E1001: **Polygenic risk score based on weight gain trajectories is predictive of childhood obesity**
Presenter: **Ana Kenney**, Pennsylvania State University, United States
Co-authors: Matthew Reimherr, Francesca Chiaromonte, Sarah Craig, Kateryna Makova

Obesity is highly heritable, yet only a small fraction of its heritability has been attributed to specific genetic variants. These variants are traditionally ascertained from genome-wide association studies (GWAS), which utilize samples with tens or hundreds of thousands of individuals for whom a single summary measurement (e.g., BMI) is collected. An alternative approach is to focus on a smaller, more deeply characterized sample in conjunction with advanced statistical models that leverage detailed phenotypes. We use novel functional data analysis (FDA) techniques to capitalize on longitudinal growth information and construct a polygenic risk score (PRS) for obesity in young children. This score is significantly higher in children with (vs. without) rapid infant weight gain. Using two independent cohorts, we show that genetic variants identified in early childhood are also informative in older children and in adults, consistent with early childhood obesity being predictive of obesity later in life. In contrast, PRSs based on SNPs identified by adult obesity GWAS are not predictive of weight gain in our cohort of children. Our research provides a strong example of a successful application of FDA to GWAS. We demonstrate that a deep, statistically sophisticated characterization of a longitudinal phenotype can provide increased statistical power to studies with relatively small sample sizes. This has the potential of shifting the existing paradigm in GWAS.

EO316 Room Bloomsbury DURATION TIME REGRESSION BEYOND THE COX MODEL Chair: Riccardo De Bin

E0929: **Link-based survival additive models with mixed types of censoring**
Presenter: **Giampiero Marra**, University College London, United Kingdom
Co-authors: Rosalba Radice, Davide Lazzaro

Existing methods for survival models are limited in that they do not often consider monotonicity constraints on the survival function, flexible covariate effects and different types of censoring mechanisms simultaneously. A methodology is discussed that addresses the three above mentioned problems by allowing for survival outcomes to be modelled using flexible parametric formulations for time-to-event data, the baseline survival function to be modelled using monotonic splines, and covariate effects to be modelled using an additive predictor incorporating several types of covariate effects. The models parameters are estimated using a carefully structured efficient and stable penalized likelihood algorithm. The proposed framework is evaluated using simulated and real data sets. The relevant numerical computations can be easily carried out using the freely available GJRM R package.

E0966: **A flexible parametric modelling framework for survival analysis**
Presenter: **Kevin Burke**, University of Limerick, Ireland

A general parametric survival modelling framework is introduced which encompasses key shapes of hazard function (constant, increasing, decreas-

ing, up-then-down, down-then-up), various common survival distributions (log-logistic, Burr type XII, Weibull, Gompertz), and includes defective distributions (cure models). This generality is achieved using four distributional parameters: two scale parameters - which, respectively, relate to accelerated failure time (AFT) and proportional hazards (PH) models - and two shape parameters. Furthermore, we advocate "multi-parameter regression" (also known as "distributional regression"), whereby more than one distributional parameter depends on covariates. In particular, we suggest introducing covariates through just one or other of the two scale parameters (covering AFT and PH models), and through a "power" shape parameter (covering more complex non-AFT/non-PH effects); the other shape parameter remains covariate-independent, and handles automatic selection of the baseline distribution. We explore inferential performance by way of simulation studies, and demonstrate the effectiveness of the framework using real data analysis.

E1053: **Copula link-based additive models for dependent right-censored event time data**
Presenter: **Robinson Dettoni**, University College London, United Kingdom
Co-authors: Giampiero Marra, Rosalba Radice

When time to event data is analysed is often assumed that the censoring mechanism is independent. This can be appropriate in many situations, in particular, when individuals are censored at the end of the study. However, in many applications, this assumption can be challenged. The aim is to introduce a class of flexible survival models in which the censoring scheme is dependent, non-informative and there are not competing risks. In particular, we show that our model is identified. Baseline functions for the event and censored times are non-parametrically estimated using monotonic P-splines. In addition, covariate effects are flexibly determined using additive predictors that allow for a vast variety of covariate effects, whereas parameter estimation is reliably carried out within a penalised maximum likelihood framework with integrated automatic multiple smoothing parameter selection. The square root (n)-consistency and asymptotic normality of the proposed flexible dependent estimator are derived. The finite sample properties of the estimators are investigated via a Monte Carlo simulation study which highlights the bias when dependent censoring is ignored and the good empirical performance of our framework. The proposal is illustrated using liver transplants data. The discussed models and methods have been implemented in the R package GJRM to allow for transparent and reproducible research.

E1078: **Estimating treatment effects in non-Markov multi-state models**
Presenter: **Jon Michael Gran**, University of Oslo, Norway

Multi-state models, as a generalization of traditional time-to-event models, is a convenient framework for analysing transitions between a possible large number of states. Estimation and covariate adjustment can be based on any traditional hazard model, separately for each transition intensity, before overall outcome measures are derived, e.g. using the Aalen-Johansen estimator. We show that some of these outcome measures, such as state transition probabilities, can be very sensitive to violation of the Markov assumption. Others, like state occupation probabilities are not. We look at two general estimation procedures for non-Markov models based on landmark subsampling, and discuss the use of Cox proportional hazard models, Aalen additive models and inverse probability of treatment weighted Nelson-Aalen estimators for causal inference versus mere prediction. The motivating example is a study on the effects of national workplace initiatives on long-term sick leave and work participation, analysing a large scale dataset linked from numerous Norwegian population-wide registries.

E1087: **First-hitting-time models for high-dimensional data: A statistical boosting approach**
Presenter: **Riccardo De Bin**, University of Oslo, Norway
Co-authors: Vegard Stikbakke

In the recent years, increasing attention has been given to first-hitting-time models, at least in the context of survival analysis. In biomedical applications, the idea is to model the health status as a stochastic process, for example a Brownian motion or a Gamma process, that degrades until it reaches a critical level (threshold), which may represent the death of a patient or the recurrence of a disease. The parameters of these processes (e.g., location and scale parameters in a Brownian motion) can depend on covariates, as well as the threshold. We develop a boosting algorithm to extend the use of first-hitting-time models to high-dimensional contexts. In particular, we focus on the situation in which low-dimensional clinical data must be combined with high-dimensional genetic data to build a prediction model. We show that the integration of these two sources of data in a first-hitting-time model is intuitive and avoids complicated weighting procedures. Finally, the novel approach is applied to a real data example.

EO492 Room G11 METHODOLOGICAL AND COMPUTATIONAL ASPECTS OF GRAPHICAL AND NETWORK MODELS Chair: Jing Zhang

E0574: **Noise injection regularization in large models with applications to neural networks and graphical models**
Presenter: **Fang Liu**, University of Notre Dame, United States

The noise injection regularization technique (NIRT) is an approach to mitigate over-fitting in large models. We will demonstrate the applications of the NIRT in two scenarios of learning large models: Neural Networks (NN) and Graphical Models (GM). For NNs, we develop a NIRT called whiteout that injects adaptive Gaussian noises during the training of NNs. We show that the optimization objective function associated with whiteout in generalized linear models has a closed-form penalty term that has connections with a wide range of regularizations and includes the bridge, lasso, ridge, and elastic net penalization as special cases; it can also be extended to offer regularizations similar to the adaptive lasso and group lasso. For GMs, we develop an AdaPtive Noisy Data Augmentation regularization (PANDA) approach to promote sparsity in estimating individual graphical models and similarity among multiple graphs through training of generalized linear models. On the algorithmic level, PANDA can be implemented in a straightforward manner by iteratively solving for MLEs without constrained optimizations. For both the NN and PANDA approaches, we use simulated and real-life data to demonstrate their applications and show their superiority or comparability with existing methods.

E0942: **Efficient estimation of change points in regime switching dynamic Markov random fields**
Presenter: **Jing Ma**, Texas A&M University, United States

Gaussian Markov random fields are commonly used to study interactions in a social or biological context. In a dynamic system, it is useful to determine when the interactions change as the underlying network evolves. We propose a method for detecting structural changes in regime switching dynamic Markov random fields, where the interactions (entries in the precision matrix) are assumed to come from two different regimes, with the transitions between the regimes modeled as linear. We introduce a fast algorithm for efficient estimation of the change points and establish oracle inequalities for the estimator. We evaluate the performance of the proposed algorithm switchNet on simulated data and apply the methodology to real data.

E0967: **Using cluster fusion regularization to estimate multiple precision matrices**
Presenter: **Brad Price**, West Virginia University, United States
Co-authors: Aaron Molstad, Ben Sherwood

A new penalized likelihood framework is discussed for estimating multiple precision matrices from different classes. This framework allows for simultaneous estimation of the precision matrices and groupings of the classes (i.e., clusters). Sparse and non-sparse estimators are proposed, both of which are solved using an iterative blockwise coordinate descent algorithm. The algorithm iterates between estimating the precision matrices given the groups and estimating the clusters given the precision matrices. Blockwise updates for computing the sparse estimator require solving an elastic net penalized precision matrix estimation problem, which we solve using a proximal gradient descent algorithm. We prove that this subalgorithm has a linear rate of convergence. In simulation studies and two real data applications, we show that our method can outperform relevant competitors which do not account for groupings of the classes, or do not account for similarity across classes.

E0999: **Sparse additive graphical models**
Presenter: **Hyonho Chun**, Boston University, United States

High-throughput technologies frequently appear in genomics, proteomics and metabolites studies, and provide ample opportunities to explore dependence among tens of thousands of biological components. Due to the complexity of biological dependence (e.g. non-linearity and presence of outliers), it is still an active research problem to estimate a compatible dependence structure and to identify subgroups from this dependence structure. We propose sparse additive graphical models by jointly estimating additive components of all variables. By estimating functional components of all variables, the estimated dependence becomes compatible to a probability density function, which is an exciting advancement in multivariate approaches as our approach yields an extra benefit of facilitating missing value imputation.

EO084 Room G3 STATISTICAL METHODS FOR RISK MANAGEMENT IN FINANCE AND INSURANCE Chair: Hideatsu Tsukahara

E0528: **Spectral backtests of forecast distributions with application to risk management**
Presenter: **Alexander Alexander John McNeil**, University of York, United Kingdom

A class of backtests for forecast distributions is studied in which the test statistic is a spectral transformation that weights exceedance events by a function of the modeled probability level. The choice of the kernel function makes explicit the user's priorities for model performance. The class of spectral backtests includes tests of unconditional coverage and tests of conditional coverage. We show how the class embeds a wide variety of backtests in the existing literature, and propose novel variants as well. The tests are illustrated by extensive examples in which we consider the performance when essential features of the forecast model are neglected, such as heavy tails and volatility.

E0758: **Detecting factors of quadratic variation in the presence of microstructure noise**
Presenter: **Daisuke Kurisu**, Tokyo Institute of Technology, Japan

A new method is developed for detecting hidden factors of Quadratic Variation (QV) of Itô semimartingales from a set of discrete observations when the market microstructure noise is present. We propose a statistical way to determine the number of factors of quadratic co-variations of asset prices based on the SIML (separating information maximum likelihood) method. In high-frequency financial data, it is important to disentangle the effects of the possible jumps and the market microstructure noise existed in financial markets. We explore the variance-covariance matrix of hidden returns of the underlying Itô semimartingales and investigate its characteristic roots and vectors of the estimated quadratic variation. We give some simulation results to see the finite sample properties of the proposed method and illustrate an empirical data analysis on the Tokyo stock market.

E1328: **Tail-risk aggregation**
Presenter: **Stefan Mittnik**, University of Munich, Germany
Co-authors: Dennis Mao

Risk aggregation is a major challenge when assessing diversified investments. Although there is ample empirical evidence that returns on financial assets correlate more strongly in down-markets and despite the growing tendency to allow for asymmetry by adopting downside-risk measures, conventional Pearson correlation still prevails in risk aggregation. We propose an alternative approach to deriving tail correlation and risk aggregation matrices associated with specific regions in joint return distributions. Specifically, we focus on tail areas to derive correlations for aggregating component risk measured in terms of expected shortfall. An empirical study illustrates that the approach can capture complex dependence structures, such as correlational asymmetry, and reliably aggregate tail risk.

E1132: **A copula approach to spatial econometrics with applications to finance**
Presenter: **Hideatsu Tsukahara**, Seijo University, Japan

Traditional models in spatial econometrics utilize a spatial weight matrix as a means to express spatial dependence, but its choice is quite arbitrary. Besides, it imposes a linear structure between dependent variables; in its simplest form, a dependent variable at one spatial unit is a linear combination of dependent variables at other spatial units. When the underlying disturbance distribution is assumed to be Gaussian or elliptical in general, the model does not allow asymmetry in dependence structure and tail dependence for spatial interactions. These restrictions are too strict in some applications, for example, in finance. Therefore, we generalize existent models to allow for some nonlinear and tail dependence in disturbance distribution by applying the copula approach which somehow reflects the spatial dependence indicated by spatial weight matrix. After discussing some properties of the resulting model, we develop an estimation method assuming a (semi)parametric copula. Simulation results illustrate the applicability of the procedure. Some real applications to financial data will be given.

E1450: **Financial systemic risk prediction with non-Gaussian orthogonal-GARCH models**
Presenter: **Patrick Walker**, University of Zurich, Switzerland
Co-authors: Marc Paolella

Several financial systemic risk indicators have been proposed after the great financial crisis with the goal of quantifying risks inherent in the markets and to anticipate future crises. One of the most popular stress indicators is based on the leading eigenvalues of the covariance matrix of a set of returns and describes the level of interconnectedness of financial assets. Originally, this stress indicator is computed from the sample covariance matrix and its dynamics are thus mainly determined by the sample size. Moreover, the sample estimator is sensitive to outliers, leading to distorted systemic risk measures. We investigate computing the risk indicator based on the forecasted conditional covariance matrices from various MGARCH models, such as CCC-, DCC- and O-GARCH. An alternative measure derived from the conditional correlation matrix is discussed. We propose a novel asymmetric, fat-tailed O-GARCH model and present an EM-algorithm for maximum likelihood estimation. Using this new robust O-GARCH model, we compute the systemic risk indicator from the implied predicted conditional correlations and achieve realistic dynamics and out-of-sample forecasts. Finally, an application to tactical asset allocation shows the economic value of the risk indicator in anticipating equity market drawdowns.

EO508 Room G5 RETROSPECTIVE SYNTHETIC CLINICAL TRIALS TO FIND NEW LIVES FOR OLD DRUGS Chair: Roy Welsch

E1973: **Harnessing machine learning, chemoproteomics, and in silico drug trials to repurpose drugs for Alzheimer's disease**
Presenter: **Mark Albers**, Harvard Medical School, United States

The exploration of transcriptomes and proteomes derived from brains with Alzheimers disease (AD) by powerful computational tools has the potential to identify pathways and targets involved in the initiation and/or progression of AD. Distinguishing primary disease drivers from secondary events is a key challenge. We use three integrated, complementary informatics approaches to discover and probe potential pathways in AD using FDA-approved drugs. First, we apply classical and network aware machine learning approaches to identify pathways and targets altered in AD brains at different stages of disease progression using data from Accelerating Medicines Partnership-AD. Second, we use systems pharmacology approaches to analyze RNA-seq and proteomic data collected from cultured human neural cells following exposure to potential disease drivers and/or FDA-approved drugs in order to discover the target selectivity of lead compounds. Moreover, these data are fed back into the predictive model as CNS-cell type-derived priors for further refinement. Third, we will introduce how in-silico drug trials in electronic health record data, which will be discussed in greater detail in this session, can evaluate candidate approved drugs using real world data. Together, these data packages will help to prioritize and design follow on clinical and translational studies to test causality of disease pathways using a repurposed drug, evaluated by positive biomarker and clinical outcomes.

E1752: **Metformin and risk of dementia incidence among 0.2 million diabetes patients: an EHR-based cohort study**
Presenter: **Bang Zheng**, Imperial College London, United Kingdom
Co-authors: Bowen Su

Type 2 diabetes and associated insulin resistance are established risk factors for dementia, a multifactorial neurodegenerative disorder, hence implying a real potential of repurposing anti-diabetes drugs for dementia prevention. We evaluated the association of metformin versus sulfonylureas prescription with the risk of dementia incidence among diabetes patients aged over 50 years, leveraging the UK Clinical Practice Research Datalink (CPRD) from 1987 to 2018. Conventional Cox regression and Propensity Score Weighting analysis were used to estimate hazard ratios (HR) and 95% confidence intervals (CI). During a median of seven years follow-up, 15,089 incident dementia cases were identified among 187,613 metformin and sulfonylureas initiators. Results from both Conventional Cox regression and Propensity Score Weighting analysis showed that metformin initiators had a lower risk of dementia compared with sulfonylureas initiators (HR= 0.91 (95% CI: 0.87-0.95) and 0.88 (95% CI: 0.81-0.96), respectively). In contrast, competing-risk analysis demonstrated the real-world cumulative dementia incidence is identical in these two groups (subdistribution HR=1.02, 95% CI: 0.98-1.06), due to higher mortality rate in sulfonylureas group. Prescription of metformin is concluded to be relatively beneficial for dementia prevention among diabetes patients, which needs to be verified in non-diabetes population.

E1950: **Repurpose anti-diabetic drugs for cancer based on causal evidence**
Presenter: **Shenbo Xu**, Massachusetts Institute of Technology, United States
Co-authors: Stan Finkelstein, Roy Welsch, Bowen Su, Bang Zheng, Marie-Laure Charpignon, Ioanna Tzoulaki

Cancer has been a worldwide health issue with increasing burden but current effective therapies for most cancer types are limited. This unmet need remains high even though tremendous investments have been made in drug R&D. As such, drug repurposing, retargeting labelled drugs for off-label diseases, has aroused more and more attention. There are 294,701 type II diabetic patients with at least one anti-diabetes prescription within Clinical Practice Research Datalink (CPRD). Among these participants, 148,983 (50.6%) individuals started by metformin monotherapy and 61,741 (21.0%) initiated sulfonylureas monotherapy. Due to analogous pretreatment clinical indications and the number of anti-diabetic drug users, we limit our cohort to metformin and sulfonylureas monotherapy initiators. Leveraging enriched EHR data with diagnosis, therapy, lab test and demographic information, we intend to repurpose anti-diabetic drugs for cancer incidence and mortality risks by emulating in-silico clinical trials among aging population based on causal evidence. In general, metformin revealed a statistically significant protective effect over sulfonylureas on cancer mortality risk, whereas their risk on cancer incidence is similar. Moreover, we have offered a systematic approach to repurpose other drugs for other diseases of interest, which demonstrate a compelling opportunity and enormous cost-saving on future cancer therapy.

E1868: **Competing risks framework for repurposing of drugs**
Presenter: **Bella Vakulenko-Lagun**, University of Haifa, Israel
Co-authors: Rebecca Betensky, Sudeshna Das, Marie-Laure Charpignon, Colin Magdamo, Yi-han Sheu, Deborah Blacker, Mark Albers

This simulation study is motivated by the research on repurposing of FDA-approved drugs for treating Alzheimers Disease (AD). Any research on AD has to account for competing death that might preclude the onset of AD. Both events, AD and death (before AD), are interrelated and happen commonly in elderly population, they cannot be assumed independent and should be addressed appropriately through a competing risks framework. We consider problems in estimation of the effects of static and assigned at baseline treatment and their measures of uncertainty. We compare several causal estimands, and address challenges in their relative interpretation. The simulation results provide insight, guidelines and practical recommendations for real life applications.

E1917: **Identifying dementia prognostic factors among diabetes individuals**
Presenter: **Aamna AlShehhi**, MIT and Khalifa University, United States

Dementia is degenerative neurodegenerative diseases. It destroys the brain normal functionality such as memories access, and decision. The failure rate of dementia therapies accounts by 99.6%. Those failures are relevant to various factors such as heterogeneous of the patients. There is increasing interest to study the cognitive impairment in Diabetes mellitus population since 60% of diabetes population are at risk of developing it. A UK population with Diabetes was identified in Clinical Practice Research Datalink. We examine dementia associated with different risk factors and comorbidities by applying statistical modeling of censored time-to-event data, namely the cox model. The cox model incorporates with inverse probability of treatment weighting using propensity score. Finally, we evaluate the model on the holdout dataset using Concordance-index and Brier scores. The results show Metformin is usually associated with a significantly lower risk of dementia in London, but not in the other regions. The results also show being male to significantly lower the risk of developing dementia in most of the cities. Also, COPD and CKD are two morbidities associated with dementia risk. The dementia risk factors and its associations vary based on the region of care. These results point to the need of adapting different care practice to different regions to account for the hidden lifestyle and environmental factors.

E1992: **Discussion for Session EO508**
Presenter: **Stan Finkelstein**, MIT, United States

The discussion will build on the set of analytical papers that were presented in this session and consider what would be the next steps, once the findings point to specific drugs that show benefits for patients with the diseases in question. There are significant regulatory and policy issues that need to be considered.

EO462 Room MAL G13 THE STEIN METHOD AND STATISTICS Chair: Robert Gaunt

E1469: **Bounds for the asymptotic distribution of the likelihood ratio**
Presenter: **Andreas Anastasiou**, University of Cyprus, Cyprus
Co-authors: Gesine Reinert

An explicit bound on the distance to chi-square for the likelihood ratio statistic is given when the data are realisations of independent and identically distributed random elements. To our knowledge, this is the first explicit bound which is available in the literature. The bound depends on the number of samples as well as on the dimension of the parameter space. We illustrate the bound with two examples: samples from an exponential distribution and samples from a normal distribution.

E1475: **To choose or not to choose a prior? That's the question!**
Presenter: **Fatemeh Ghaderinezhad**, Gent university, Belgium

A challenging question in Bayesian statistics is how choosing the prior can affect the posterior distribution. How can the posteriors derived under different priors be similar as nowadays more and more data are collected? One of the newest instruments to answer this question is Stein's method. This crafty method gives the lower and upper bounds on the Wasserstein distance, at a fixed sample size, between two posteriors resulting from different priors (even improper priors). In addition, this method plays an important role in practice. It often occurs that practitioners hesitate between two proposed priors in a given situation. This methodology then allows them to know how different the two priors actually are, and to decide whether or not it is relevant to consider both priors or just stick to one of them. In particular, when hesitating between a simple, closed-form prior and a much more complicated prior. The applicability of this method is shown by conducting some simulation studies.

E0964: **On asymptotic normality in estimation after a group sequential trial**
Presenter: **Ben Berckmoes**, University of Antwerp, Belgium
Co-authors: Anna Ivanova, Geert Molenberghs

It is proven that in many realistic cases, the ordinary sample mean after a group sequential trial is asymptotically normal if the maximal number of observations increases. We derive that it is often safe to use naive confidence intervals for the mean of the collected observations, based on the ordinary sample mean. Our theoretical findings are confirmed by a simulation study. Links with Stein's method are explored.

E0745: **Pseudo-binomial Approximation to $(k1,k2)$-runs**
Presenter: **Amit Kumar**, Indian Institute of Technology Bombay, India
Co-authors: Neelesh Shankar Upadhye

It is known that the distribution of $(k1,k2)$-runs, arising from non-identical Bernoulli trials, is intractable. Hence, it is important to approximate $(k1,k2)$-runs by a suitable distribution, with reasonable accuracy. We discuss pseudo-binomial approximation to $(k1,k2)$-runs using Stein's method. The approximation results are of optimal order and improve upon the existing results in the literature.

E0786: **Approximations related to sum of m-dependent random variables**
Presenter: **Neelesh Shankar Upadhye**, Indian Institute of Technology Madras, India
Co-authors: Amit Kumar, Palaniappan Vellaisamy

The sum of m-dependent random variables concentrated on $\mathbb{Z}_+ = \{0,1,2,\dots\}$ is considered. We propose to approximate it with power series distributions and obtain the error bounds using Stein's method. We also discuss some relevant results for power series distributions such as Stein operator, uniform and non-uniform bounds for Stein magic factor, and etc. As special cases, we discuss two applications, namely, 2-runs and (k_1,k_2)-runs and compare the bound with the existing bounds.

EO290 Room MAL G14 COMPUTATIONAL STATISTICS IN DISTRIBUTION THEORY Chair: Andriette Bekker

E0397: **Circular mean variance mixture models**
Presenter: **Priyanka Nagar**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

One of the widely used methods for introducing asymmetry into a model is by means of a mean mixture approach. Many wrapped circular models have been proposed based on this method. However, the need for flexible circular models is still prevalent. A new general class of flexible spherical models is introduced based on mean variance mixture modeling. Special cases of this new class are studied in detail. Given the complex structure of this class, an EM algorithm based approach for performing maximum likelihood estimation is considered. The practicality of the proposed distribution is illustrated through a data application.

E0837: **Geometric Polya-Aeppli process**
Presenter: **Leda Minkova**, Sofia University, Bulgaria

A new point process, called Geometric Polya-Aeppli process, is introduced with underlying exponential distribution. We provide the system of differential equations for the distribution of the number of events of the Geometric Polya-Aeppli process up to time t and discuss some of its properties. The new process is an extension of the well-known Polya-Aeppli process, as well as the standard geometric process with underlying exponential distribution.

E1152: **Inference of the weighted type I and type II bivariate Polya-Aeppli distributions**
Presenter: **Claire Geldenhuys**, University of Pretoria, South Africa
Co-authors: Rene Ehlers

The purpose is to discuss and compare the method of moments estimation of the model parameters for the existing and proposed weighted Type I and Type II bivariate Polya-Aeppli distributions. Two different weight functions are used for the weighted distributions. We also compare the success rates of the estimation method for the different distributions and consider scenarios where some of the distributions might be preferable over the others. Real and synthetic data are used to compare the fit of the existing and proposed weighted Type I and Type II bivariate Polya-Aeppli models to data.

E1162: **Weighted type I and type II bivariate Polya-Aeppli distributions**
Presenter: **Rene Ehlers**, University of Pretoria, South Africa
Co-authors: Claire Geldenhuys

The theory of the Type I and Type II bivariate Polya-Aeppli distributions is extended to the weighted case by using two different weight functions. This allows for more flexible distributions that can be fitted to bivariate count-valued data. Some properties of the distributions are discussed. The dispersion of the new distributions compared to the existing distributions is also discussed by using the bivariate Fisher index of dispersion.

E1269: **Bayesian estimation of a generalised entropy functional using alternative Dirichlet priors**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Andriette Bekker

Entropy (of which Shannon's is arguably most popular) is a common and widely studied measurement of information contained within a system. Entropy is most often defined as a functional of a probability structure, and the practical problem of estimating entropy from (sometimes small) samples in many applied settings remains a challenging and relevant problem. Previously unconsidered Dirichlet generators are introduced as possible priors for an underlying countably discrete model (in particular, the multinomial model). Resultant estimators for the generalised entropy $H(p)$, which include popular entropy choices, under the considered priors and assuming squared error loss are derived and studied. Particular cases of these proposed priors will be of interest, and their effect on the estimation of the generalised entropy subject to different parameter scenarios will be investigated.

EO655 Room MAL G15 RECENT DEVELOPMENTS IN BAYESIAN CAUSAL INFERENCE Chair: Genevieve Lefebvre

E0690: **Prior constraints in estimation of causal effects in natural experiments**
Presenter: **Sara Geneletti**, London School of Economics, United Kingdom
Co-authors: Gianluca Baio, Jose Pina-Sanchez, Aidan OKeeffe, Sylvia Richardson, Federico Ricciardi, John Paul Gosling

Topics are considered where constraints are placed on prior distributions in order to obtain causal effect estimates. The first topic is the estimation of the causal effect of statins (a type of cholesterol lowering drug) in the UK population using a regression discontinuity design. We considered both continuous and binary outcomes and imposed constraints on the prior distributions of some parameters in order to stabilise and obtain causal effect estimates. The second topic involves generating continuous values for the severity of non-custodial sentences. A long-standing issue in criminology is that sentence types come in two flavours – custodial sentences measured in days and non-custodial sentences measured as factor levels – making it difficult to compare the two types of outcomes and evaluate the effect of policy changes on sentencing. We describe a method to extend a continuous severity score based on sentence length to non-custodial outcomes. This method involves using prior constraints to impose an ordering by ensuring the severity of non-custodial outcomes cannot exceed certain thresholds. The data thus generated can be used as part of an interrupted time series design to estimate the causal effects of changing sentencing guidelines.

E0969: **Priors comparison in Bayesian mediation framework with binary outcome**
Presenter: **Anne Philippe**, Universite de Nantes, France

Co-authors: Jean-Michel Galharret

Mediation refers to a causal phenomenon in which the effect of an exposure variable X on an outcome Y can be decomposed into a direct effect and an indirect effect via a third variable M. We propose a Bayesian analysis to estimate these effects. We show that including information into the prior distribution helps to improve the quality of the estimation. The proposed informative prior is based on the knowledge of previous observations on the same population (e.g. longitudinal study) or from an independent study done under similar conditions. One of the usual issue in the mediation analysis is to test the existence of the direct and the indirect effect. We propose a testing procedure based on credible intervals and its asymptotic properties. Though simulation, we compare this procedure with the tests used in mediation analysis. Finally, we apply our approach to real data from a longitudinal study on the well-being of children in school.

E0980: **Modeling baseline treatment effects in Bayesian network meta-analysis of disconnected networks**

Presenter: **Audrey Beliveau**, University of Waterloo, Canada

Co-authors: Sergiu Pocol

Network meta-analysis is a set of statistical tools conventionally used to establish comparative efficacy/safety of more than two interventions using data extracted from a systematic literature review of randomized controlled trials. In recent years, Bayesian implementations of network meta-analysis have been prominent since they lend themselves naturally into ranking of treatment efficacy and into health-economic decision modeling. Disconnected networks arise when some of the treatments of interest are not compared head-to-head within a study nor indirectly through studies with treatments in common. The analysis of disconnected networks is typically frowned upon due to the lack of a trusted gold-standard method. The standard contrast-based model for the network meta-analysis of connected networks treats baseline treatment effects as fixed but treating these as random makes the estimation of contrasts between disconnected treatments possible. Using a normal distribution on the baseline treatment effects inherently assumes that the baseline treatment effects are exchangeable across studies and that a normal distribution represents accurately the variation across studies. We explore empirically to what extent this assumption of normal and exchangeable baseline treatment effects could be a problem in real applications and whether alternative distributions such as the Student t distribution could help mitigate problems.

E1121: **Bayesian analysis of correlated exposure biomarkers subject to a limit of detection**

Presenter: **Lawrence McCandless**, Simon Fraser University, Canada

Biomarkers are widely used in perinatal epidemiology to examine the health effects of environmental chemical exposures during pregnancy. However a difficulty with biomarkers is that the chemical concentrations at low doses are often left-censored by a limit of detection. Further complicating the analysis, chemical biomarkers are often highly correlated because the exposure occur in mixtures and they have the same underlying exposure source (eg. polychlorinated biphenyl mixtures). The pattern of missing data and correlated predictor variables complicates efforts to measure the causal effects of individual and combined exposure to multiple chemical agents on reproductive and child health outcomes. We describe a novel Bayesian approach to examine the role of correlation between chemical biomarkers to enhance imputation of missing data, and we use this information to improve estimation of the health effects of low level chemical exposures.

E1204: **Assessing causal effects in a longitudinal observational study with truncated outcomes and nonignorable missing data**

Presenter: **Alessandra Mattei**, University of Florence, Italy

Co-authors: Michela Bia, Andrea Mercatanti

Important statistical issues pervade the evaluation of effects of training programs for unemployed people. In particular the fact that offered wages are observed and well-defined only for subjects who are employed (truncation by death), and the problem that information on the employment status and wage can be lost over time (attrition) raise methodological challenges for causal inference. We present an extended framework for simultaneously addressing the aforementioned problems, and thus answering important substantive research questions in training evaluation observational studies with covariates, a binary treatment and longitudinal information on employment status and wage affected by the presence of missing data. There are two key features of this framework: we use principal stratification to properly define the causal effects of interest and we adopt a Bayesian approach for inference. The proposed framework allows us to partially answer an open issue in economics: the assessment of the trend of reservation wage over the duration of unemployment. We apply our framework to evaluate causal effects of foreign language training programs in Luxembourg, using administrative data on the labor force (IGSS-ADEM dataset). Our findings might be an incentive for the employment agencies to better design and implement future language training programs.

EO076 Room MAL G16 DEPENDENCE MODELS AND COPULAS Chair: Fabrizio Durante

E1209: **Predicting extreme surges from sparse data using a copula-based hierarchical Bayesian spatial model**

Presenter: **Jonathan Jalbert**, Polytechnique Montraal, Canada

Co-authors: Melina Mailhot, Christian Genest, Nicholas Beck

A hierarchical Bayesian model is proposed to quantify the magnitude of extreme surges on the Atlantic Coast of Canada with limited data. Generalized extreme-value distributions are fitted to surges derived from water levels measured at 21 buoys along the coast. The parameters of these distributions are linked together through a Gaussian field whose mean and variance are driven by atmospheric sea-level pressure and the distance between stations, respectively. This allows for information sharing across the original stations and for interpolation anywhere along the coast. The use of a copula at the data level of the hierarchy further accounts for the dependence between locations, allowing for inference beyond a site-by-site basis. It is shown how the extreme surges derived from the model can be combined with the tidal process to predict potentially catastrophic water levels.

E1715: **Dependence properties and Bayesian inference for asymmetric multivariate copulas**

Presenter: **Marta Crispino**, University of Oslo, Norway

Co-authors: Stephane Girard, Julyan Arbel

Some new theoretical properties of a broad class of asymmetric copulas will be introduced. Such copulas are obtained as a combination of multiple, usually symmetric, copulas. We will also focus on a subclass of Liebscher copulas obtained by combining comonotonic copulas which are characterized by an arbitrary number of singular components. Furthermore, we will introduce a novel iterative representation for general Liebscher copulas which de facto insures uniform margins, thus relaxing a constraint of the original construction. This iterative construction proves useful for inference by developing an Approximate Bayesian computation sampling scheme. The inferential procedure will be demonstrated on simulated data.

E0430: **Copula-based analysis of multivariate dependence patterns between dimensions of poverty in Europe**

Presenter: **Cesar Garcia Gomez**, Universidad de Valladolid, Spain

Co-authors: Ana Perez Espartero, Mercedes Prieto-Alaiz

It is widely recognised that poverty is a multidimensional phenomenon involving not only income, but also other aspects such as education or health. In this multidimensional setting, analysing the dependence between dimensions becomes an important issue, since a high degree of dependence could exacerbate poverty. We propose measuring the multivariate dependence between the dimensions of poverty in Europe using copula-based methods. This approach focuses on the positions of individuals across dimensions, allowing for other types of dependence beyond linear correlation. In particular, we analyse how orthant dependence between the dimensions of the AROPE rate has evolved in the EU-28 countries between 2008 and 2014 by applying non-parametric estimates of multivariate copula-based generalisations of Spearman's rank correlation coefficient. We find a

general increase in the dependence between dimensions, regardless of the coefficient used. Moreover, countries with higher AROPE rates also tend to experiment more dependence between its dimensions.

E0821: **Quantifying and estimating asymmetric dependence**
Presenter: **Florian Griessenberger**, University Salzburg, Austria
Co-authors: Wolfgang Trutschnig
Standard dependence measures considered in the literature like Pearson correlation, Spearman rank correlation or Schweitzer and Wolff's σ are symmetric, i.e. they assign each pair of random variables (X,Y) the same dependence as they assign the pair (Y,X). Since dependence structures are in general not symmetric (in contrast to independence, which is a symmetric concept), the classical dependence measures fail to detect asymmetry. The recently developed R-package qad (short for quantification of asymmetric dependence) aims at detecting asymmetries in samples. It estimates the dependence of the second variable on the first one and vice versa, and additionally quantifies the asymmetry of the underlying dependence structure. The main objectives are to sketch the idea underlying the copula-based dependence measure, to present the most relevant mathematical properties of the underlying estimator and to illustrate its capabilities by some examples.

E1030: **Geometric structure in dependence models and applications**
Presenter: **Elisa Perrone**, University of Massachusetts Lowell, United States
The geometric properties of copulas are explored in order to address dependence modeling challenges in several applications, such as hydrology and finance. In particular, we study the class of discrete copulas, i.e., restrictions of copulas on uniform grid domains, which admits representations as convex polytopes. First, we give a geometric characterization of discrete copulas with desirable stochastic constraints in terms of the properties of their associated convex polytopes. In doing so, we draw connections to the popular Birkhoff polytopes, thereby unifying and extending results from both the statistics and the discrete geometry literature. Then, we further consolidate the statistics/discrete geometry bridge by showing the significance of our geometric findings to (1) construct entropy-copula models useful in hydrology, and (2) design test statistics for stochastic monotonicity properties of interest in finance. Finally, we discuss extension to analyze discrete copulas with positive dependence constraints, such as total positivity.

EO612 Room CLO 101 ADVANCES IN HIGH-DIMENSIONAL STATISTICS Chair: Rajen D Shah

E0281: **Tuning parameter calibration for large and high-dimensional data**
Presenter: **Johannes Lederer**, Ruhr-University Bochum, Germany
Some aspects of tuning parameter calibration for high-dimensional estimators are discussed.

E0848: **Analysis of networks via the sparse beta-model**
Presenter: **Chenlei Leng**, Warwick, United Kingdom
Co-authors: Mingli Chen, Kengo Kato
The Sparse beta-Model (SβM) is proposed, a new network model that interpolates the celebrated Erdos-Renyi model and the more recent beta-model by distinguishing global and local sparsity. The model has attractive statistical and computational properties for sparse networks with degree heterogeneity. Numerical results are presented to illustrate its usage.

E0909: **Regression with high-dimensional categorical data using nonconvex penalties**
Presenter: **Benjamin G Stokell**, University of Cambridge, United Kingdom
Co-authors: Ryan J Tibshirani, Rajen D Shah
Categorical data arise in a number of application areas, often with large numbers of levels. We propose a method for estimation in linear models with such covariates. Our method is called 'SCOPE', standing for Sparse Concave Ordering and Penalisation Estimator. Within each categorical variable, coefficients are ordered and their adjacent differences penalised by a concave function. It can quickly be computed exactly using a dynamic programming algorithm, exploiting the separable structure of the optimisation objective. We study its theoretical properties and give conditions under which the oracle property holds. This approach can also be used to fit logistic regression models.

E1043: **Covariance structure induced by sparsity in non-standard domains**
Presenter: **Heather Battey**, Imperial College London and Princeton University, United Kingdom
An estimate of a covariance or inverse covariance matrix is an essential ingredient in many multivariate statistical procedures. When the dimension of the matrix is large relative to the sample size, the sample covariance matrix is inconsistent in the matrix norms relevant for applications and its non-invertibilty renders many techniques in multivariate analysis infeasible. Structural assumptions are necessary in order to restrain the estimation error, even if this comes at the expense of some approximation error if the structural assumptions fail to hold. We will discuss the non-sparse structure induced on the original space of covariance matrices by imposing sparsity in the matrix logarithmic domain. We will also show the converse result that any covariance matrix possessing such structure is logarithmically sparse. Generalisations of this structure will then be discussed.

E1106: **Spectral deconfounding**
Presenter: **Domagoj Cevid**, ETH Zurich, Switzerland
Co-authors: Peter Buehlmann, Nicolai Meinshausen
High-dimensional regression methods which rely on the sparsity of the ground truth, such as the Lasso, might break down in the presence of confounding variables. If a latent variable affects both the response and the predictors, the correlation between them changes. Such hidden confounding can be represented as a high-dimensional linear model where the sparse coefficient vector is perturbed. For this model, we develop and investigate a class of methods that are based on running the Lasso on preprocessed data. The preprocessing step consists of applying certain spectral transformations that change the singular values of the design matrix. We show that, under some assumptions, one can achieve the optimal ℓ_1-error rate for estimating the underlying sparse coefficient vector and illustrate the performance on a genomic dataset.

EO090 Room CLO 102 SHRINKAGE METHODS FOR LARGE TIME SERIES MODELS Chair: Ines Wilms

E0549: **Sparse change-point VAR models**
Presenter: **Arnaud Dufays**, Namur University, Belgium
Co-authors: Jeroen Rombouts, Yong Song, Zhuo Li
Change-point (CP) VAR models face a dimensionality curse due to the proliferation of parameters that arises when new segments are detected. To handle large data set, we introduce the Sparse CP VAR process that determines which parameters truly vary when a break is detected. By doing so, the number of new parameters to estimate at each segment is drastically reduced and the CP dynamic is easier to interpret. The Sparse CP VAR model disentangles the dynamics of the mean parameters and the covariance matrix. The latter is driven by an infinite hidden Markov framework while the former stands for a CP dynamic with shrinkage prior distributions on the first-difference parameters. We argue that limiting the number of possible breaks in the mean parameters has several theoretical and empirical advantages over the standard practice of time-varying parameter (TVP) models. An exhaustive Monte Carlo study highlights that the framework operates for detecting the correct number of breaks per model parameter in small and medium-sized dimensional settings. The Sparse CP VAR model also takes advantage of common breaks in the cross-sectional dimension to more precisely estimate them. Our applications on 7 and 25 macroeconomic systems highlight that the Sparse CP

VAR model help to interpret the detected breaks. In addition, the Sparse CP VAR model outperforms several recent TVP and CP VAR models in terms of log predictive densities.

E0570: **Sparse single-equation error correction models in high dimensions**
Presenter: **Etienne Wijler**, Maastricht University, Netherlands
Co-authors: Stephan Smeekes

In this paper we propose the Single-equation Penalized Error Correction Selector (SPECS) as an automated estimation procedure for dynamic single-equation models with a large number of potentially (co)integrated variables. By extending the classical single-equation error correction model, SPECS enables the researcher to model large cointegrated datasets without necessitating any form of pre-testing for the order of integration or cointegrating rank. We show that SPECS is able to consistently estimate a linear combination of the cointegrating vectors, while simultaneously enabling the correct recovery of sparsity patterns in the parameter space. The results are derived in an asymptotic framework where both the number of (ir)relevant variables as well as the number of observations diverge. Novel eigenvalue conditions with broad applicability to penalized regression estimators are presented.

E0632: **Detecting changes in the covariance structure of high dimensional time series using random matrix theory**
Presenter: **Sean Ryan**, Lancaster University, United Kingdom
Co-authors: Rebecca Killick

A novel method is proposed for detecting changes in the covariance structure of high dimensional time series. The approach uses a test statistic that is based on the "natural" distance between two covariance matrices. As a result, the performance of this geometrically inspired method does not depend on the underlying structure of the covariance matrix. We demonstrate that this compares favourably with other approaches. Using results from Random Matrix Theory we explore the asymptotic behaviour of our method in the high dimensional setting (i.e. the number of variables is a comparable to the length of the data). We also explore the finite sample performance of our method via a range of simulations.

E0830: **Desparsified lasso in time series**
Presenter: **Robert Adamek**, Maastricht University, Netherlands
Co-authors: Ines Wilms, Stephan Smeekes

The Desparsified Lasso is a high-dimensional estimation method which provides uniformly valid inference. We extend this method to a time series setting under mixingale assumptions allowing for non-Gaussian, serially correlated and heteroskedastic processes, where the number of regressors can possibly grow faster than the time dimension. We first derive an oracle inequality for the (regular) Lasso, relaxing the commonly made exact sparsity assumption to a weaker alternative, which permits many small but non-zero coefficients. The weak sparsity coupled with the mixingale assumption means this inequality can also be applied to the (inherently misspecified) nodewise regressions performed in the Desparsified Lasso. This allows us to establish the uniform asymptotic normality of the Desparsified Lasso under general conditions. Additionally, we show consistency of a long-run variance estimator, thus providing a complete set of tools for performing inference in high-dimensional linear models. Finally, we perform a simulation exercise to demonstrate the small sample properties of the Desparsified Lasso in common time series settings.

E1182: **Penalized graphical models for cross-sectional and longitudinal microbiome data**
Presenter: **Christian L. Mueller**, Simons Foundation, United States
Co-authors: Zachary Kurtz

Learning statistical association networks from microbial targeted amplicon sequencing (TAS) data holds the promise to unravel the organizational structure of microbes in their natural habitats, ranging from human gut to large-scale marine ecosystems. TAS count data are compositional (or relative abundance) data due to experimental limitations, thus requiring dedicated statistical methods for network inference. For cross-sectional microbiome data, a popular statistical framework is based on log-ratio data transformations from the field of compositional data analysis, followed by shrinkage-based graphical model inference. We extend this framework to longitudinal and time series data using a latent variable approach. For short longitudinal data sets we model the autocorrelations in the samples via structured low-rank components, followed by shrinkage-based graphical model inference. For sufficiently long time series, we first employ a Bayesian non-parametric approach that takes into account the compositionality of the time series, and then employ shrinkage-based latent graphical models on the residuals. We illustrate the efficacy of these approaches on several longitudinal and time series data sets from the human gut microbiome.

EO298 Room Court ISSUES IN CONTEMPORARY CLUSTERING Chair: Taps Maiti

E0515: **On model-based learning and directional outlier detection**
Presenter: **Cristina Tortora**, San Jose State University, United States

Classification can be lucidly defined as the process of assigning group labels to sets of observations. When a finite mixture model is used for classification in either an unsupervised, semi-supervised, or supervised setting, one can refer to this process as model-based learning. We will present a paradigm for parameterizing contamination and skewness within variants of the mixtures of shifted asymmetric Laplace (SAL) distributions. These models will be able to provide both group labels for like observations and detect whether an observation is an outlying point, unifying the fields of model-based learning and outlier detection. Of particular interest are the multiple scaled variants of the mixtures of SAL distributions which allow for directional contamination and skewness, resulting in contours that do not have the traditional elliptical shapes. Explicit details regarding the development of the proposed models will be provided and an expectation-maximization based parameter estimation scheme will be outlined. The classification performance of these models will be demonstrated using simulated and real data sets.

E0839: **A semiparametric Bayesian model for biclustering**
Presenter: **Alejandro Murua**, University of Montreal, Canada
Co-authors: Fernando Quintana

Motivated by classes of problems frequently encountered in the analysis of gene expression data, we propose a semiparametric Bayesian model to detect biclusters, that is, subsets of individuals sharing similar patterns over a set of conditions. The approach is based on the well-known plaid model. By assuming a truncated stick-breaking prior we also find the number of biclusters present in the data as part of the inference. Evidence from a simulation study shows that the model is capable of correctly detecting biclusters and performs well compared to some competing approaches. The flexibility of the proposed prior is demonstrated with applications to the analysis of gene expression data (continuous responses) and histone modifications data (count responses).

E1092: **Effect of penalisation on a mixture of factor analysers**
Presenter: **Nam-Hwui Kim**, University of Waterloo, Canada
Co-authors: Ryan Browne

Factor analysers can be used to obtain a parsimonious estimate of component-wise covariance matrices in a finite mixture model. In addition, one could achieve further parsimony in estimated covariance matrix by penalising on the factor loading matrix. However, an increasing magnitude of penalisation coefficient may result in degenerate factor loading estimates, which may have an adverse effect on maximum likelihood estimation of model parameters. To this end, we investigate the effect of penalisation on sparse estimation of parameters in a finite mixture of factor analysers. We also investigate the effect of such estimates in model-based clustering settings.

E1381: **Optimal transport, mean partition, and uncertainty assessment in cluster analysis**
Presenter: **Beomseok Seo**, The Pennsylvania State University, United States
Co-authors: Jia Li, Lin Lin
In scientific data analysis, clusters identified computationally often substantiate existing hypotheses or motivate new ones. Yet the combinatorial nature of the clustering result, which is a partition rather than a set of parameters or a function, blurs notions of mean, and variance. This intrinsic difficulty hinders the development of methods to improve clustering by aggregation or to assess the uncertainty of clusters generated. We overcome that barrier by aligning clusters via optimal transport. Equipped with this technique, we propose a new algorithm to enhance clustering by any baseline method using bootstrap samples. Cluster alignment enables us to quantify variation in the clustering result at the levels of both overall partitions and individual clusters. Set relationships between clusters such as one to one match, split, and merge can be revealed. A covering point set for each cluster, a concept kin to the confidence interval, is proposed. The tools we have developed will help address the crucial question of whether any cluster is an intrinsic or spurious pattern. Experimental results on both simulated and real data sets are provided. The corresponding R package OTclust is available on CRAN.

E1471: **Model-based learning via mixtures of contaminated shifted asymmetric Laplace distributions**
Presenter: **Brian Franczak**, MacEwan University, Canada
Classification can be lucidly defined as the process of assigning group labels to sets of observations. When a finite mixture model is used for classification in either an unsupervised, semi-supervised, or supervised setting, one can refer to this process as model-based learning. We will discuss a mixture of shifted asymmetric Laplace (SAL) distributions and extensions thereof. Specifically, we will focus on the development of mixtures of contaminated shifted asymmetric Laplace factor analyzers (MCSALFA). Compared to the well-known mixtures of Gaussian distributions, the mixtures of SAL distributions can parameterize skewness in addition to both location and scale, making it well suited for the analysis of data with homogeneous subpopulations that are not symmetric. In addition to providing a classification of similar observations, the MCSALFA will also provide a classification of an observation as being either 'good' or 'bad', unifying the fields of classification and outlier detection. Details regarding the development of the MCSALFA will be provided and an alternating-expectation conditional-maximization based parameter estimation scheme will be outlined. The classification performance of these mixtures will be demonstrated using several real data sets.

EO468 Room Jessel STRUCTURAL CHANGES IN MULTIVARIATE AND HIGH-DIMENSIONAL DATA Chair: Piotr Fryzlewicz

E0626: **Asymptotically distribution-free change-point detection for multivariate and non-Euclidean data**
Presenter: **Lynna Chu**, Iowa State University, United States
Co-authors: Hao Chen
The focus is on testing and estimation of change-points, locations where the distribution abruptly changes, in a sequence of multivariate or non-Euclidean observations. While the change-point problem has been extensively studied for low-dimensional data, advances in data collection technology have produced data sequences of increasing volume and complexity. Motivated by the challenges of modern data, we study a non-parametric framework that can be effectively applied to various data types as long as an informative similarity measure on the sample space can be defined. The existing approach along this line has low power and/or biased estimates for change-points under some common scenarios. To address these problems, we present new tests based on similarity information that exhibit substantial improvements in detecting and estimating change-points. In addition, under some mild conditions, the new test statistics are asymptotically distribution free under the null hypothesis of no change. Analytic p-value approximation formulas to the significance of the new test statistics are derived, making the new approaches easy off-the-shelf tools for large datasets. The effectiveness of the new approaches are illustrated in an analysis of New York taxi data.

E0824: **Identifying coordinates with change in high-dimensional panel data using tail-summed scores**
Presenter: **Anica Kostic**, London School of Economics and Political Science, United Kingdom
Co-authors: Piotr Fryzlewicz
Detection of possibly aligned change-points is considered in the high-dimensional mean-shift model. The interest is not only in the detection of the number and locations of change-points, but also in determining which panel components have undergone a change at a given time. As changes in some components can be insignificant when considered individually, we propose a new iterative multiple testing and signal discovery procedure, which considers components in groups and uses tail-summed scores. We show its attractive properties both in theory and in practice, with particular advantages over the state of the art when the mean changes are small but dense across the panel.

E0974: **A block segmentation scheme for structural break detection in large scale high-dimensional non-stationary VAR models**
Presenter: **Abolfazl Safikhani**, University of Florida, United States
Co-authors: George Michailidis
Many real time series data sets exhibit structural changes over time. A popular model for capturing their temporal dependence is that of Vector Autoregressions (VAR), which can accommodate structural changes through time evolving transition matrices. The problem then becomes to both estimate the (unknown) number of structural break points, together with the VAR model parameters. An additional challenge emerges in the presence of very large data sets, namely on how to accomplish these two objectives in a computational efficient manner. In this paper, we propose a novel procedure which leverages a block segmentation scheme (BSS) that reduces the number of model parameters to be estimated through a regularized least squares criterion. Specifically, BSS examines appropriately defined blocks of the available data, which when combined with a fused lasso based estimation criterion, leads to significant computational gains without compromising on the statistical accuracy in identifying the number and location of the structural breaks. The procedure is scalable to large high-dimensional time series data sets with a computational complexity proportional to square root of sample size. Extensive numerical work on synthetic data supports the theoretical findings and illustrates the attractive properties of the procedure. Finally, an application to a neuroscience data set exhibits its usefulness in applications.

E1114: **Inference of break-points in high-dimensional time series**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Weining Wang, Wei Biao Wu
A new procedure is considered for detecting structural breaks in mean for high-dimensional time series. We target breaks happening at unknown time points and locations. In particular, at a fixed time point our method is concerned with either the biggest break in one location or aggregating simultaneous breaks over multiple locations. We allow for both big or small sized breaks, so that we can 1), stamp the dates and the locations of the breaks, 2), estimate the break sizes and 3), make inference on the break sizes as well as the break dates. Our theoretical setup incorporates both temporal and cross-sectional dependence, and is suitable for heavy-tailed innovations. We derive the asymptotic distribution for the sizes of the breaks by extending the existing powerful theory on local linear kernel estimation and high dimensional Gaussian approximation to allow for trend stationary time series with jumps. A robust long-run covariance matrix estimation is proposed, which can be of independent interest. An application on detecting structural changes of the US unemployment rate is considered to illustrate the usefulness of our method.

E1470: **Covariance change point detection and identification with high-dimensional functional data**
Presenter: **Ping-Shou Zhong**, University of Illinois at Chicago, United States
Co-authors: Shawn Santo
High-dimensional functional data appear in practice when a dense number of repeated measurements is taken on a large number of variables for a

relatively small number of experimental units. The spatial temporal dependence and high-dimensional nature of the data structure make statistical analysis a challenge. A procedure is developed to detect and identify change points among covariance matrices from high-dimensional functional data. A new test statistics is proposed for change point detection whose asymptotic distribution is established under mild assumptions. We further estimate the locations of the change points if exist. The estimator is proven to be consistent under a set of mild conditions. Its rate of convergence depends on the data dimension, sample size, number of repeated measurements, and signal-to-noise ratio. Computation efficiency is carefully addressed to cope with the large number of repeated measurements and variables measured. Simulation results demonstrate that the size of the test is well controlled at the nominal level, and the locations of multiple change points can accurately be identified. A functional neuroimaging data set is demonstrated to identify points of change in functional connectivity of the human brain.

EO372 Room MAL 152 ESTIMATION AND HYPOTHESIS TESTING FOR DEPENDENT STOCHASTIC PROCESSES Chair: Salim Bouzebda

E0279: **Uniform in bandwidth convergence rate of the kernel regression estimator adaptive to intrinsic dimension**
Presenter: **Thouria El Hadjali**, Universite de Technologie de Compiegne, France
Co-authors: Salim Bouzebda, Boutheina Nemouchi

The focus is on the uniform in bandwidth consistency of kernel-type regression estimators of the regression function $\mathbb{E}(\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x})$ derived by modern empirical process theory, under weaker conditions on the kernel than previously used in the literature. Our theorems allow data-driven local bandwidths for these statistics. We extend existing uniform bounds on kernel regression estimator and making it adaptive to the intrinsic dimension of the underlying distribution of $\mathbf{X}$ which will be characterizing by the so-called intrinsic dimension. Moreover, we show, in the same context, the uniform in bandwidth consistency for nonparametric inverse probability of censoring weighted (I.P.C.W.) estimators of the regression function under random censorship.

E0331: **A dependent stochastic process in Bayesian nonparametrics**
Presenter: **Mame Diarra Fall**, University of Orleans and CNRS, France

Dirichlet processes (DP) and Dirichlet processes mixtures (DPM) have emerged as cornerstones in Bayesian nonparametric models. The former can be used as a prior on a probability mass function, while the latter is a suitable prior for a probability density function. The focus is on a special case of dependent Dirichlet process mixtures (DDPM). We show how this can be used to handle challenging problems in image reconstruction.

E0406: **Nonparametric regression for locally stationary functional data**
Presenter: **Aboubacar Amiri**, Charles de Gaulle University, France

The problem of the nonparametric regression of a real random variable on a non-stationary time series of functional data is addressed. We focus on the estimation of the regression function using a kernel approach. We introduce an estimator of the regression operator that takes into account the non-stationary behavior of the data-generating process. The mean square error and the almost sure convergence of the proposed estimator are derived. In addition, a central limit theorem on the regression estimator is established. Asymptotic results are established with convergence rates, whereas the asymptotic constants are explicitly calculated by assuming that the covariate is a local stationary and strong mixing functional process.

E0443: **A space-time process for extremes: Application to precipitation data**
Presenter: **Gwladys Toulemonde**, Universita de Montpellier, France
Co-authors: Jean-Noel Bacro, Carlo Gaetan, Thomas Opitz

The statistical modeling of space-time extremes in environmental applications is a valuable approach to understand complex dependences in observed data and to generate realistic scenarios for impact models. Motivated by hourly rainfall data in Southern France presenting asymptotic independence, we propose a novel hierarchical model for high threshold exceedances defined over continuous space and time by embedding a space-time Gamma process convolution for the rate of an exponential variable, leading to asymptotic independence in space and time. This construction permits keeping marginal distributions which are coherent with univariate extreme value theory. Statistical inference is based on a pairwise likelihood for the observed censored excesses. The practical usefulness of our model is illustrated on the previously-mentioned hourly precipitation data set and comparisons with alternative censored Gaussian random fields are discussed.

E0490: **Modified portmanteau tests for ARMA models with dependent errors**
Presenter: **Yacouba Boubacar Mainassara**, Universite Bourgogne Franche-Comte, France

The purpose is to derive the asymptotic distribution of residual autocovariances and autocorrelations of autoregressive moving-average (ARMA) models with uncorrelated but non-independent innovations. We then deduce modified portmanteau statistics and we establish the asymptotic distribution of the proposed statistics.

EO474 Room MAL 153 COMPUTATION AND LIKELIHOOD IN BIOSTATISTICAL AND ENVIRONMENTAL MODELS Chair: Anuradha Roy

E0272: **Log concave densities in symbolic data analysis**
Presenter: **Carlo Drago**, University of Rome Niccolo Cusano, Italy

Big data require the extraction of relevant information from the data separating them from the noise. In this sense, symbolic data analysis allow us to work efficiently with large data sets by providing relevant approaches which can be used in order to extract the relevant information from data. Symbolic data allows us to represent and considering explicitly the uncertainty present on the data which can be not considered by using some aggregations as the mean. Various proposals already exists as intervals, boxplots, histograms, densities, beanplots, mixtures, and so on. We propose a new approach based on a symbolic data based on log-concave densities. These computations and representations recently attracted many interest. Log-concave densities has interesting properties and they can be considered as symbolic data when it could be clear the effect of some groups of observations or some outliers on the estimated density. Another advantage is that they do not need to choose relevant parameters as the bandwidth. Another relevant point is that there are various approaches to estimate the log-concave density. In this sense, the log-concave density estimation seems to be very useful and appropriated in various fields of application like environmental problems. We will show the features of this approach theoretically, by simulation and on a real application.

E0647: **Mixtures of factor analyzers with covariates for clustering multiply censored dependent variables**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan

Censored data arise frequently in diverse applications in which observations to be measured may be subject to some upper and lower detection limits due to the restriction of experimental apparatus such that they are not exactly quantifiable. Mixtures of factor analyzers with censored data (MFAC) have been recently proposed for model-based density estimation and clustering of high-dimensional data under the presence of censored observations. We consider an extended version of MFAC with covariates to accommodate multiply censored dependent variables and develop two analytically feasible EM-type algorithms for computing maximum likelihood estimates of the parameters with closed-form expressions. Moreover, we provide an information-based method to compute asymptotic standard errors of mixing proportions and regression coefficients. The utility and performance of our proposed methodologies are illustrated through simulated data and two real data examples.

E0730: **A new all-purpose generic transformation with applications in multivariate modelling and missing value imputation**
Presenter: **Ravindra Khattree**, Oakland University, United States

Copulas have been used in various applications in biomedical sciences and finance. We suggest copulas as the generic all-purpose transformations

which can enable one to apply various standard multivariate procedures more efficiently and with better statistical properties and results. More specifically, we consider the problem of transformation of any continuous data to multivariate normality using copulas as a device for defining the transformation. Such a transformation effectively enables us to model a variety of problems involving non-normal data using the classical multivariate statistical techniques. We evaluate and illustrate various applications where analyses using the appropriate copula transformations result in substantial improvement in implementation, interpretation, prediction as well as in the corresponding models. Finally, we use this approach for multiple imputation problem for the missing data when the underlying distribution is nonnormal.

E0903: **Testing multiple means in two-level compound symmetry multivariate data**
Presenter: **Ivan Zezula**, P.J. Safarik University, Slovakia
Co-authors: Daniel Klein
A simple multivariate model $X = M + E$ is considered, where M is a location or mean matrix, and E is an error matrix. The variance matrix of E is assumed to be doubly-exchangeable, i.e. block-wise compound symmetry. We develop several test procedures for multiple testing of possibly different mean matrices sharing the same variance matrix. Properties of different tests are compared via simulations.

E1354: **Overcoming challenges in the analysis of longitudinal discrete data**
Presenter: **Justine Shults**, Perelman School of Medicine at University of Pennsylvania and Children's Hospital of Philadelphia, United States
Challenges are faced in the analysis of longitudinal discrete data that we do not face with continuous outcomes. Likelihoods for discrete outcomes can be complex, especially for longitudinal data with serial correlation and overdispersion. Semi-parametric approaches such as generalized estimating equations (GEE) are appealing because they do not require specification of the full likelihood. However, it is possible to unknowingly obtain estimates for which there is no valid parent distribution. We demonstrate some of the challenges we face in the analysis of longitudinal discrete data and show how we can overcome some of the difficulties via the first-order Markov maximum likelihood based approach.

EO667 Room Senate COMMUNICATING STATISTICS AND DATA SCIENCE TO THE MASSES Chair: Jennifer Green

E1128: **Communicating controversial statistical results to the public**
Presenter: **Jessica Utts**, University of California, Irvine, United States
One of the fundamental aspects of statistical studies is dealing with uncertainty. But most members of the public want concrete information and are not comfortable with uncertainty. Communicating uncertainty is even more important (and difficult) when the topic is controversial. The speaker was the co-author of a report for the U.S. Congress assessing the statistical evidence for psychic abilities in U.S. government-sponsored research. The report received widespread attention and the speaker was interviewed for various media formats including print, radio, television, and documentary films. This experience provided both the opportunities and responsibilities for communicating controversial statistical information to a broad lay public. It also highlighted the challenges of providing statistical information in short sound bites and on live television in an interview format. This talk will discuss that experience, and offer suggestions for statisticians who wish to communicate results to the masses.

E1212: **Clickbait, fake news, and accidental misleads**
Presenter: **Liberty Vittert**, Olin Business School, United States
Data Science (/statistics) has become an integral part of the news cycle and the world we live in. How do we communicate statistics to the public in the world of 20 second sound bites, "fake news", and certainty. Statistics by definition measures uncertainty, but when told to be "sure" where does the grey-line lie between being statistically correct and invited to come back on TV (ever). From receiving highly negative comments from other statisticians (sometimes fairly) to receiving praise from the news organizations, where is the balance in communicating statistics in the media?

E1272: **Writing for newspapers, magazines, and comics**
Presenter: **Regina Nuzzo**, American Statistical Association, United States
Some of my experiences writing about statistical concepts for publications such as Nature News, New Scientist, and the New York Times will be highlighted. I will also discuss a recent project in which I teamed up with an award-winning comics artist to do a feature-length comic about statistical issues in forensic science. I will discuss ideas for incorporating lay-language writing into the curriculum, workplace, and professional organizations.

E1319: **If data-driven journalism is the future, what do journalists need to know about statistics**
Presenter: **Trevor Butterworth**, Sense About Science USA, United States
In 2010, Sir Tim Berners-Lee, inventor of the world-wide web said "journalists need to be data-savvy" because "data-driven journalism is the future." Was he right? And if so, what do journalists need to understand about statistics and how can statisticians engage the media. This talk will report on three projects by Sense About Science USA that have engages journalists and statisticians: STATScheck, a free service for journalists staffed by volunteer statisticians that has addressed hundreds of queries on how to understand data on deadline; STATS workshops for newsrooms, journalism schools, and journalism organizations; and communication workshops for statisticians run in conjunction with the American Statistical Association. What have we learned from these projects? What needs to be done to improve statistical literacy in the media?

E1600: **You have been asked to talk to the media, now what?**
Presenter: **Kendra Schmid**, University of Nebraska Medical Center, United States
As a statistician, you probably never expect to get asked for media interviews, especially related to your own work. What happens when you do? This talk will discuss the presenters experiences with different types of media (newspapers, magazines, television, talk shows, live demonstrations, media tours), ways in which one might get contacted by media, processes from initial contact to finish, and tips for positively representing yourself, your work, and your institution through these interactions. Additionally, this talk will discuss aspects of working with companies and presenting research results through media, including the fine line between scientific work and product endorsement.

EO322 Room CLO 203 RECENT DEVELOPMENT IN EXPERIMENTAL DESIGNS AND INDUSTRIAL STATISTICS Chair: Chang-Yun Lin

E1350: **Representation of multivariate Bernoulli distributions with a given set of specified moments**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy
Co-authors: Patrizia Semeraro
A new but simple method is proposed to characterise multivariate Bernoulli variables belonging to a given class, i.e., with some specified moments. Within a given class, this characterisation allows us to easily generate a sample of mass functions. It also provides the bounds that all the moments must satisfy to be compatible and the possibility to choose the best distribution according to a certain criterion. For the special case of the Frechet class of the multivariate Bernoulli distributions with given margins we find a polynomial characterization of the class. Our characterization allows us to have bounds for the higher order moments. An algorithm and its use in some examples is shown. Possible connections with design of experiments will be illustrated.

E1388: **Two-level block designs under model uncertainty**
Presenter: **Pi-Wen Tsai**, National Taiwan Normal University, Taiwan
Co-authors: Steven Gilmour
Two-level designs are widely used for screening experiments with the goal to identify the few active factors which have major effects. Blocking is

a common technique when there is systematic variation among experimental units. Most work on two-level block designs focuses on the method of replacement for finding the best blocking scheme of regular and irregular factorial designs, and we are forced to sacrifice the estimations of one or more factorial effects where all main effects are orthogonal to blocks. We discuss a model-robust block criterion which respects experimenters' prior knowledge on the importance of each effect. It is shown that several minimum aberration criteria for block designs are limiting cases of this model-robust block criterion. Additionally, a coordinate-exchange algorithm is developed to generate new classes of block designs under model uncertainty. We will demonstrate that by relaxing the requirement of orthogonal blocking, more appropriate designs can be recommended under different experimenters priors.

E1391: **Prediction properties of optimum response surface designs**
Presenter: **Steven Gilmour**, KCL, United Kingdom
Co-authors: Luzia Trinca, Heloisa Oliveira, Cesar Oliveira

Prediction capability is considered an important issue in response surface methodology. Following the line of argument that a design should have several desirable properties we have extended an existing compound design criterion to include prediction properties. Prediction of responses and of differences in response are considered. Point and interval predictions are allowed for. Extensions of existing graphical tools for inspecting prediction performances of the designs in the whole region of experimentation are also introduced. The methods are illustrated with two examples.

E1404: **Optimal non-collapsing space-filling designs for irregular experimental regions**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan

Space-filling and non-collapsing are two important properties in designing computer experiments. We study how the non-collapsing, space-filling designs for irregular experimental regions can be generated efficiently by the proposed metaheuristic methods. We solve this optimal design problem using variants of the discrete particle swarm optimization (DPSO) approaches. Numerical results, including an application in data center thermal management, are used to illustrate the performances of the proposed algorithms. Based on these numerical results, we assert that the most efficient approach is to reformulate the target optimal design problem as a constrained optimization problem and then use a modified DPSO to solve the constrained optimization problem.

E1437: **Construction, properties, and analysis of group-orthogonal supersaturated designs**
Presenter: **Ryan Lekivetz**, SAS Institute Inc., United States

A new method is introduced for constructing supersaturated designs that is based on the Kronecker product of two carefully-chosen matrices. The construction method leads to a partitioning of the columns of the design such that the columns within a group are correlated to the others within the same group, but are orthogonal to any factor in any other group. The resulting designs are referred to as group orthogonal supersaturated designs (GOSSDs). This group structure is leveraged to obtain an unbiased estimate of the error variance and to develop an effective, design-based model selection procedure. The designs can also be used in group screening.

EO114 Room CLO 204 NOVEL STATISTICAL METHODS AND APPLICATIONS FOR MEDICAL DATA Chair: Mihye Ahn

E1861: **A dimension reduction method for group analysis of functional neuroimaging data**
Presenter: **Mihye Ahn**, University of Nevada Reno, United States

Recently, much attention has been paid to the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. A simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors than to some existing methods. Finally, we apply the proposed method to ADNI data.

E1753: **Age related network efficiency and the role of multi-modal neural underpinning**
Presenter: **Seonjoo Lee**, Columbia University/New York State Psychiatric Institute, United States

The neuroanatomical underpinnings of cognitive reserve have been investigated using either structural magnetic resonance imaging (MRI) or functional MRI. A multimodal data-fusion method is applied to a set of structural MRI, DTI and FLAIR data acquired from 167 cognitive normal participants in order to elucidate associations between aging and brain network efficiency. We compared the performance with the results when all modalities are analyzed separately.

E1944: **On simulating ultra high-dimensional multivariate data**
Presenter: **Alfred Schissler**, University of Nevada, Reno, United States
Co-authors: Alex Knudson

In this era of Big Data, it is critical to realistically simulate data to conduct informative Monte Carlo studies. This is often problematic when data are inherently multivariate while at the same time are (ultra-) high dimensional. This situation appears frequently in observational data found on online and in high-throughput biomedical experiments (e.g., RNA-sequencing). Due to the difficulty in simulating realistic correlated data points, researchers often resort to simulation designs that posit independence — greatly diminishing the insight into the empirical operating characteristics of any proposed methodology. A major challenge lies in the computational complexity involved in simulating these massive multivariate constructions. We motivate and propose a fairly general, scalable procedure to simulate high-dimensional multivariate distributions with pre-specified marginal characteristics and a covariance matrix. We apply our method to simulate RNA-sequencing data sets with heterogeneous negative binomial marginals.

E1601: **Evaluation of biomarkers for treatment selection using individual participant data from multiple clinical trials**
Presenter: **Chae Ryon Kang**, University of Pittsburgh, United States
Co-authors: Holly Janes, Parvin Tajik, Henk Groen, Ben Mol

Biomarkers that predict treatment effects may be used to guide treatment decisions, thus improving patient outcomes. A meta-analysis of individual participant data (IPD) is potentially more powerful than a single-study data analysis in evaluating markers for treatment selection. The motivation comes from the IPD that were collected from two randomized controlled trials of hypertension and pre-eclampsia among pregnant women to evaluate the effect of labor induction over expectant management of the pregnancy in preventing progression to severe maternal disease. The existing literature on statistical methods for biomarker evaluation in IPD meta-analysis have evaluated a markers performance in terms of its ability to predict risk of disease outcome, which do not directly apply to the treatment selection problem. We propose a statistical framework for evaluating a marker for treatment selection given IPD from a small number of individual clinical trials. We derive marker-based treatment rules by minimizing the average expected outcome across studies. The application of the proposed methods to the IPD from two studies in women with hypertension in pregnancy is presented.

E1959: **Graphical models for data integration and mediation analysis**
Presenter: **Min Jin Ha**, UT MD Anderson Cancer Center, United States
Co-authors: Veerabhadran Baladandayuthapani, Francesco Stingo

Integrative network modeling of data arising from multiple genomic platforms provides insight into the holistic picture of the interactive system, as well as the flow of information across many disease domains. The basic data structure consists of a sequence of hierarchically ordered datasets for each individual subject, which facilitates integration of diverse inputs, such as genomic, transcriptomic, and proteomic data. A primary analytical task in such contexts is to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers, dictated by multiple platforms, and exhibit both undirected and directed relationships. We propose a multi-layered Gaussian graphical model (mlGGM) to investigate conditional independence structures in such multi-level genomic networks. We use a Bayesian node-wise selection approach that coherently accounts for the multiple types of dependencies in mlGGM, that is used for finding causal factors for outcome variables via mediation analysis.

EO556 Room MAL 253 UNCERTAINTY IN WEATHER, CLIMATE, AND HYDROLOGICAL FORECASTS Chair: Annette Moeller

E0284: **Ensemble methods for weather prediction**

Presenter: **Roberto Buizza**, Scuola Superiore Sant'Anna (SSSA) Pisa, Italy

Ensemble methods based on a limited number of numerical integrations, have so far proven to be the only feasible way to estimate the probability distribution function (PDF) of forecast states. Probabilistic forecast products are generated by computing statistics based on a finite number of members, which are generated to produce accurate and reliable forecasts. The implementation of operational ensembles in 1992 followed years of research in predictability, which saw many scientists both in academia and in operational numerical weather prediction (NWP) centers investigating how best to deal with the sources of forecast uncertainties. The operational implementations at the European Center for Medium-Range Weather Forecasts (ECMWF, Europe) and at the National Centers for Environmental Prediction (NCEP, US) induced a paradigm shift in NWP from providing a single forecast, to issuing a range of forecasts that can be used to identify possible future scenarios, compute the probability of events of interest, and in general to estimate forecast confidence levels. Ensembles helped the development of subseasonal and seasonal prediction systems. In climate studies, ensembles are used to estimate the range of possible future scenario. In NWP, today, ensembles are used also to estimate the PDF of initial states. A brief overview will be provided about the key characteristics of ensemble methods used in weather prediction, to estimate the initial and forecast PDF.

E0671: **Statistical postprocessing under model and climate changes: A challenge**

Presenter: **Stephane Vannitsem**, Free University of Brussels and Royal meteorological institute of Belgium, Belgium

Co-authors: Jonathan Demaeyer

One important hypothesis behind post-processing of weather forecasts and climate projections is the stationarity of the processes under investigations. Model changes are however very frequent and nature is also experiencing climate changes. These modifications are affecting the quality of corrected forecasts and projections. The statistical and dynamical properties of bias correction and linear post-processing are investigated when the system under interest is affected by model errors and is experiencing parameter modifications in the context of simple well controlled systems. Challenges and way to cope with these modifications are discussed.

E0234: **Using wavelets to verify the correlation structure of meteorological forecast fields**

Presenter: **Sebastian Buschow**, University of Bonn, Germany

Co-authors: Petra Friederichs

When predicting meteorological fields at high spatial resolutions, one important aspect of forecast quality is their spatial correlation structure. While current weather models and post-processing techniques may be unable to foresee the precise timing and location of small-scale event, a fair verification should nonetheless reward their progress in representing the overall spatial pattern. Since naive point-wise approaches fail to reward highly resolved forecasts in the presence of displacement errors, numerous so-called spatial verification techniques have emerged. The purpose is to deal with a recently developed methodology for the spatial verification of deterministic, as well as ensemble forecasts, based on discrete wavelet transformations: By projecting observed and predicted fields on a new set of basis functions with varying spatial scale, orientation and location, we can estimate a local wavelet spectrum at each grid point. These spectra intuitively summarize the distribution of spatial variability across scales and directions and can (under appropriate assumptions) directly be related to the spatial covariances themselves. We will briefly introduce the relevant theoretical background before demonstrating how the wavelet-approach can be used to analyse and compare precipitation structures in high-resolution ensemble forecasts over Germany.

E0327: **Statistical post-processing of water level forecasts**

Presenter: **Sandor Baran**, Faculty of Informatics, University of Debrecen, Hungary

Co-authors: Stephan Hemri, Mehrez El Ayari

Accurate and reliable probabilistic forecasts of hydrological quantities like runoff or water level are beneficial to various areas of society. Probabilistic state-of-the-art hydrological ensemble prediction models are usually driven with meteorological ensemble forecasts. Hence, biases and dispersion errors of the meteorological forecasts cascade down to the hydrological predictions and add to the errors of the hydrological models. The systematic parts of these errors can be reduced by applying statistical post-processing. For a sound estimation of predictive uncertainty and an optimal correction of systematic errors, statistical post-processing methods should be tailored to the particular forecast variable at hand. Former studies have shown that it can make sense to treat hydrological quantities as bounded variables. For flexible post-processing of multi-model ensemble forecasts of water level a doubly truncated Bayesian model averaging (BMA) method is introduced, which generalizes the truncated normal BMA model for wind speed calibration. BMA weights and model parameters are estimated with the help of the EM algorithm for truncated normal mixtures. A case study based on water level data for gauge Kaub of river Rhine reveals a good predictive skill of doubly truncated BMA compared both with the raw ensemble and the reference ensemble model output statistics approach.

E0393: **Spatially consistent postprocessing of probabilistic cloud cover forecasts**

Presenter: **Stephan Hemri**, MeteoSwiss, Switzerland

Co-authors: Christoph Spirig, Jonas Bhend, Jan Rajczak, Lionel Moret, Mark Liniger

Even though numerical weather prediction models (NWPs) are run at increasingly high resolutions, raw ensemble forecasts still tend to be biased and underdispersed. Hence, statistical postprocessing is expected to improve forecast skill and to provide a more reliable estimation of forecast uncertainty. At MeteoSwiss a project on statistical postprocesssing of ensemble forecasts for spatial fields of different variables has recently been launched, integrating the available NWP ensemble forecasts and observations. Predictions of cloud cover are impaired by Switzerland's complex terrain and the high frequency of low stratus over the Swiss Plateau, which is poorly represented in the raw model predictions. First tests based on ensemble model output statistics and analog based approaches showed an increase in univariate forecast skill. While the application of empirical copula based approaches like ensemble copula coupling or the Schaake shuffle turned out to provide promising baseline scenarios, the methods to generate physically realistic forecast scenarios still need to be improved. A consistent representation of the spatial structure of cloud cover and realistic forecast scenarios are desired for graphical forecast products serving the general public. Possible methods and first results will be discussed to overcome this challenge.

EC812 Room MAL 254 CONTRIBUTIONS IN SPATIAL STATISTICS Chair: Anastassia Baxevani

E1639: **Restricted spatial regression methods: Implications for inference**
Presenter: **Kori Khan**, The Ohio State University, United States
Co-authors: Catherine Calder

The issue of spatial confounding between the spatial random effect and the fixed effects in regression analyses has been identified as a concern in the statistical literature. Multiple authors have offered perspectives on this issue and potential solutions. For the areal spatial data setting, we show that many of the methods designed to alleviate spatial confounding can be viewed as special cases of a general class of models. Extending terminology currently in use, we refer to this class as Restricted Spatial Regression (RSR) models. Using this insight, we offer a mathematically based exploration of the impact that RSR methods have on inference for regression coefficients for the linear model. We show that the use of these methods have counterintuitive consequences which defy the general expectations in the literature. In particular, our results and accompanying simulations suggest that RSR methods will typically perform worse than non-spatial methods. A simulation study of count data indicates these results may extend to the generalized linear model setting.

E1642: **Supervised dimension reduction for spatial data**
Presenter: **Christoph Muehlmann**, Technical University of Vienna, Austria
Co-authors: Klaus Nordhausen, Hannu Oja

In regression tasks a higher number of predictors makes modeling very demanding and increases the computational cost significantly. Supervised dimension reduction (SDR) addresses these issues by reducing the number of predictors prior building the actual model. Sliced inverse regression (SIR) is one popular SDR method that is well established for iid data. SIR was recently also extended to the time series case. However, there seem not to be any SDR methods for spatial data. In the spatial data context, it is natural to assume that measurements that are closer together show more similarity than measurements taken far apart. Similarly, in spatial regression the response variable maybe not only depending on the on-site predictors but also on predictors in the vicinity. There are many regression models considering spatial dependence but issues with a high number of predictors are still remaining. We extend SIR to spatial data recorded on a grid by formulating it in a blind source separation model to extract a subspace of the neighboring predictors that carries the most information of the response variable. Furthermore, practical guidelines on how to choose the dimension of the subspace as well as spatial lags of interest are given.

E1034: **Re-examining the similarity threshold of Andresen's spatial point pattern S-Index**
Presenter: **Inger Fabris-Rotelli**, University of Pretoria, South Africa
Co-authors: Rene Kirsten, Greg Breetske

Andresen's S-index is used widely by geographers, specifically in criminology literature. The S-index represents the percentage of spatial units that have similar spatial patterns in both point patterns and ranges from 0 to 1. The test is subjective in that it delineates spatial similarity and dissimilarity at $S = 0.8$ and $S = 0.5$ respectively. We propose a technique to remove this subjectively by using the second-order nature of the spatial data. Am improved, stronger concluding test is thus set up.

E1739: **Non-stationary wrapped Gaussian processes with sparse precision matrices**
Presenter: **Isa Marques**, University of Goettingen, Germany
Co-authors: Thomas Kneib, Nadja Klein

Directional data, i.e., data consisting of angles, can be found across many areas of science, such ecology, biology, environmental sciences, or medicine. The special nature of such data means that conventional methods for linear data are not suitable. Nonetheless, few attempts have been made to develop flexible models for periodic data, namely spatial models. Gaussian random fields are one of the most important building blocks for hierarchical models for spatial data. Yet the need to factorize dense covariance matrices renders them quite computationally expensive. We introduce a spatial model for wrapped Gaussian data which, given the empirical equivalence between Gaussian and Gaussian Markov random fields, considerably reduces computational complexity by using sparse matrix algorithms. The selection of appropriate hyperpriors for the Gaussian fields' parameters is an important and sensible topic, specially given that these are not consistently estimable from a single realization. Consequently, we develop penalized complexity priors for the model's hyperparameters that are practically useful and tunable. The posterior distribution is assessed with an adaptive Markov chain Monte Carlo. Finally, we extend previously existing directional data models by allowing for covariates in both the mean and the covariance structure, as well as for a nugget effect.

E0208: **Robust cross-variogram estimators and their distributions**
Presenter: **Alfonso Garcia-Perez**, UNED, Spain

Let $Z(s) = (Z_1(s),...,Z_p(s))^t$ be a multivariate spatial process that satisfies the intrinsic stationarity property. Assuming that we have a sample of $Z(s)$ at n locations, we measure the statistical association between the random components of Z with the correlation coefficient and the spatial dependence with the variograms. To capture the association both within components of $Z(s)$ and across s, we need the cross-variogram, defined for collocated data, i.e., assuming that each location has all variables Z_i measured, as $2\gamma_{ij}(h) = E[(Z_i(s+h) - Z_i(s))(Z_j(s+h) - Z_j(s))]$. We define robust estimators of the cross-variogram and we obtain a saddlepoint approximation for their sample distributions, assuming a multivariate scale contaminated normal distribution as model.

EC799 Room MAL 354 CONTRIBUTIONS IN STATISTICAL MODELLING II Chair: Inmaculada Barranco-Chamorro

E1567: **A new regression model for discrete data allowing for overdispersion**
Presenter: **Roberto Ascari**, University of Milano-Bicocca, Italy
Co-authors: Sonia Migliorati, Eduardo Buono

Binomial regression is commonly used for modeling discrete data which represent the number of successes in a fixed number of independent trials. This approach is very popular, but is inadequate in case of overdispersion, i.e. when real data show a larger variance than the one assumed by the binomial distribution. This excess of variability is typically due to violation of the i.i.d. assumption of the binary variates forming the binomial outcome. A possible way of dealing with overdispersion is to compound the binomial with a distribution defined on the unit-interval. If the beta distribution is chosen, the beta-binomial is attained. We define a new distribution, the flexible beta-binomial (FBB), obtained compounding the binomial with the flexible beta. The FBB can be expressed as a finite mixture of beta-binomial distributions. Further, we define a GLM-type regression model based on the FBB distribution and show that thanks to its parametrization, it allows for a form of the variance and the intraclass correlation coefficient easily interpretable in terms of overdispersion. Inferential issues are dealt with a Bayesian approach through a Hamiltonian Monte Carlo algorithm.

E0185: **Modelling income distribution using the log Student t distribution: New evidence for EU countries**
Presenter: **Mercedes Prieto-Alaiz**, University of Valladolid, Spain
Co-authors: Carmelo Garcia Perez, Francisco Javier Callealta Barroso

The ability of the three-parameter log Student t distribution to model the size distribution of income is studied. Its theoretical properties and the economic interpretation of its parameters are analysed and some theoretical results on the distribution characteristics are obtained. The model is fit to income data for EU25 and several years and its goodness of fit measures are compared with those of the other three parameter distributions that have been considered as successful in modelling income distributions. It is concluded that the log Student t distribution is the best fitting in the vast

majority of the countries and years considered. One of the reasons for these good fits is the possibility of reproducing the effect of a mode in the low-income environment.

E1762: **Regularized Babington-Smith ranking model**
Presenter: **Jong-june Jeon**, University of Seoul, Korea, South

The aim is to present the regularization method for the Babington-Smith model that derives academically important ranking models. By regularizing parameters of the model, we construct a continuum class of ranking models that bridges Babington-Smith model and Bradley-Terry-Mallows model. Through the regularization we can account for an unusual characteristic such as intransitivity of preferences in the model. We also propose the computational algorithm based on constrastive divergence to estimate the parameters in our model and prove its convergence property of the algorithm. We investigate the model selection problem in our model and provide the results of numerical simulations.

E0315: **On generalized marginal inhomogeneity model for multidimensional contingency tables**
Presenter: **Kengo Fujisawa**, Tokyo University of Science, Japan
Co-authors: Kouji Tahata

For the analysis of contingency tables with ordinal categories, the interest is whether each marginal distribution is homogeneous or not. It may be more appropriate to apply a certain marginal inhomogeneity model when the marginal homogeneity model does not fit. The generalized marginal inhomogeneity model is proposed by using a continuous cumulative distribution function. Using this model, we prove that the marginal homogeneity model is decomposed into two models. We investigate the model-fitting by numerical simulations.

E1866: **New nonparametric approach to correct response bias on ordinal categorical data using Anchoring vignette**
Presenter: **Mariko Takagishi**, Osaka university, Japan

In questionnaire surveys, Likert type questions (e.g., 1-strongly agree,...,5-strongly disagree) are often observed. However, since how to interpret each category is different among respondents, direct comparison among respondents is not straightforward. We call the bias that occurs by ignoring this interpersonal incomparability problem as "response bias". Anchoring vignette is a tool to correct response bias in observed response. In the Anchoring vignette framework, a method for correcting the bias is applied to the questionnaire data before the statistical data analysis. Several existing correction methods are proposed, such as an ordinal regression based, and Item response theory based. The possible question would be "which correction method should I use before the data analysis?". So far this problem has not been discussed well, because for the existing methods, the "corrected value" is defined in various ways and thus cannot compare the property among them. Therefore, we introduce a new comprehensive statistical modeling for correction which includes the existing correction methods as special cases, and propose a new simple nonparametric correction method based on the new statistical model. In addition, we derive the property of the proposed corrected value which is useful for data analysis.

EC807 Room MAL 355 CONTRIBUTIONS IN MULTIVARIATE STATISTICS Chair: Ioulia Papageorgiou

E0194: **The MLE-3D algorithm for the 2-fold growth curve model**
Presenter: **Joseph Nzabanita**, University of Rwanda, Rwanda
Co-authors: Dietrich von Rosen, Martin Singull

There is a growing interest in the analysis of multi-way data. In many studies the inference about the dependencies in three-way data is done using the third order tensor normal model, where the focus is on the estimation of the variance-covariance matrix which has a Kronecker product structure. Little attention is paid to the structure of the mean, though, there is a potential to improve the analysis by assuming a structured mean. Assuming a trilinear structure for the mean in the tensor normal model, a 2-fold growth curve model is formulated and a maximum likelihood estimation based algorithm for estimating parameters is proposed. Simulation studies show that the proposed algorithm performs well.

E1852: **Structural equation modeling considering cluster structure**
Presenter: **Ippei Takasawa**, Doshisha University, Japan
Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

Structural equation modeling (SEM) is a method that clarifies the relationships between observed variables and latent factors using confirmatory factor analysis. Generalized structured component analysis (GSCA) is one type of SEM and is a component-based approach like PCA and is estimated using alternating least squares (ALS). GSCA can express many flexible models because it consists of three structural equations: from variables to variables, from variables to components, and vice versa. To consider the heterogeneity of data, fuzzy clusterwise GSCA (FCGSCA) is proposed as extended GSCA. FCGSCA is a method such as simultaneously estimating cluster labels for subjects and these path components by each cluster. It is difficult to assume a different path diagram for each cluster because forms of path diagrams are decided before estimation. Accordingly, FCGSCA uses same path diagram and each cluster feature is interpreted as differences of coefficients. However, there are situations in which path coefficients for each cluster are not very different although the data is heterogeneous and detection of clusters with different path coefficients is needed. Therefore, we propose a method that estimates each path diagram to be more different and makes it easy to interpret each cluster feature by constraints of the coefficient matrix.

E1652: **Three-mode PCA for finding a solution intermediate between Tucker3 and Parafac**
Presenter: **Aya Nakashima**, Osaka University, Japan
Co-authors: Kohei Adachi

Three-mode PCA (3MPCA) refers to the modified PCA procedures specially designed for analyzing a three-mode data array. In 3MPCA, the data array is approximately decomposed into three loading matrices and a core array. This array describes the relationships among the components occupying the columns of the three loading matrices. Tucker3 and Parafac models are among popular 3MPCA ones. The Tucker3 model can be considered as too less restrictive, in that its core array is unconstrained. Thus, it is not easy to interpret the array. In contrast, the Parafac model is too restrictive, as the elements in its core array are forced to be zeros except the super-diagonal elements. Thus, Parafac likely provides the solutions with bad fit to data. Those discussions suggest that the model is useful which is intermediate between Tucker3 and Parafac, i.e., whose core array includes a suitable number of zero elements. For exploring such an intermediate model, we propose a new 3MPCA procedure. In this procedure, the Tucker3 loss function is minimized subject to the constraint that a specified number of core elements are exactly zeros, with which elements are zeros being unknown. Therefore, the optimal locations of the zero elements and nonzero parameter values are to be estimated simultaneously. For the estimation, we present an alternating least squares algorithm. Its behaviors are assessed in a simulation study and the proposed procedure is illustrated with real data examples.

E1879: **Asymptotic distribution of the linear discriminant function with two-step monotone sample**
Presenter: **Nobumichi Shutoh**, Kobe University, Japan

The aim is to derive the asymptotic distribution of the linear discriminant function constructed by the estimators on the basis of monotone sample. Because the estimators of the parameters have more complicated forms as the number of missing patterns k increases, we show the results for the case of $k = 2$ for simplicity. The main results can be applied to approximate the misclassification probabilities for the linear discriminant function. Monte Carlo simulation is conducted in order to evaluate the accuracy of the approximation.

E1858: **Mixture of factor analyzers for NMAR missing data**
Presenter: **Yuki Morioka**, Doshisha University Graduate School of Culture and Information Science, Japan

Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

Probabilistic principal component analysis (PPCA) is a statistical multivariate method, widely used for research in several fields, that operates with the assumption of continuous latent variables. In PPCA, we can tackle missing values by using an EM algorithm. To deal with data with heterogeneous structures, PPCA was extended to a mixture model. To improve clustering accuracy, the PPCA mixture model is also extended to a mixture of factor analyzers, which can deal with observed errors of all variables are different structures. However, this method does not assume a missing data mechanism of NMAR. If the missing data mechanism is NMAR, the result of existing method is influenced by the missing values. Therefore, we proposed a new mixture of factor analyzers. We can obtain the parameters of our proposed method using an EM algorithm. We demonstrate the validity of the proposed method through a numerical example.

EG267 Room MAL 251 CONTRIBUTIONS ON STATISTICS FOR INSURANCE AND ACTUARIAL SCIENCES Chair: Udi Makov

E0859: **Why does the human die: Cohort-wise mortality prediction under the survival energy hypothesis**

Presenter: **Yasutaka Shimizu**, Waseda University, Japan

A quite new methodology is proposed for the mortality prediction. We assume there exists the survival energy in human beings: we are born with a certain initial quantity of survival energy, and the energy changes stochastically in time. The human dies if the energy decreases and hits the level zero, that is, the time of death is the first hitting time of the survival energy processes. We suppose the survival energy follows a diffusion process and define the mortality as the first hitting time distribution of the process. We estimate the unknown parameters in the stochastic differential equations via the least squares estimation by fitting the model to the empirical hitting time distribution function using the historical data from "Human Mortality Database". We shall illustrate by the real data analysis that such a "structural approach" to the human energy gives a surprisingly good prediction for long-future's mortality.

E1574: **The Poisson-Lognormal regression model for mean and dispersion with an application to insurance ratemaking**

Presenter: **Natalia Hong**, London School of Economics, United Kingdom

Co-authors: Ryan Ho, George Tzougas

Within the actuarial field, the family of mixed Poisson models has been used extensively to model claim count data. The main focus is to present an extension of the Poisson-lognormal (PLN) regression model where both the mean and the dispersion parameters of the distribution are modelled as a function of explanatory variables. The adopted framework allows us to capture the stylized characteristics of the data in a more complete way. We propose a quite simple Expectation-Maximization type algorithm for maximum likelihood estimation of the model. Finally, a real data application using motor insurance data is examined and both the a priori and a posteriori, or Bonus-Malus, premium rates resulting from the model are compared to those determined by the Negative Binomial Type I and the Poisson-Inverse Gaussian regression models with regression structures on every parameter.

E1895: **Some moment-indeterminate distributions from actuarial science**

Presenter: **Christian Kleiber**, Universitaet Basel, Switzerland

The moment problem asks whether or not a given distribution is uniquely determined by its moments. The existence of moment-indeterminate distributions has been known since the work of Stieltjes in the late 19th century; the most widely known example appears to be the lognormal distribution. Recent research has shown that the phenomenon is more widespread than previously thought. Specifically, distributions that are not determined by their moments arise in the modelling of size distributions in economics and related fields. We study determinacy issues for certain claim size (aka claim severity) distributions occurring in non-life insurance.

E1931: **A general family of mixed exponential models applied to heavy-tailed losses**

Presenter: **George Tzougas**, London School of Economics and Political Science, United Kingdom

Co-authors: Dimitris Karlis

Regression modelling involving heavy-tailed response distributions, which have heavier tails than the exponential distribution, has become more and more popular in many insurance settings including non-life insurance. Mixed Exponential models can be considered as a natural choice for the distribution of heavy-tailed claim sizes since their tails are not exponentially bounded. However, apart from very few cases, such as the traditional Pareto regression model, this family of models has not been studied in depth. The main reason is that mixed Exponential models are not usually tractable because their likelihood is complicated and hence its maximization needs a special effort. The aim is to introduce a general family of mixed Exponential regression models with varying dispersion which can efficiently capture the tail behavior of losses. Our main achievement is that we present an Expectation-Maximization (EM) type algorithm which can facilitate maximum likelihood (ML) estimation for our class of mixed Exponential models which allows for regression specifications for both the mean and dispersion parameters. Finally, a real data application based on motor insurance data is given to illustrate the versatility of the proposed EM type algorithm.

E1569: **Optimal design of long term care insurance for Medicaid seniors using a multi-state Markov model of mortality/morbidity**

Presenter: **Colin Ramsay**, University of Nebraska-Lincoln, United States

Co-authors: Victor Oguledo

Consider a retired U.S. senior (i.e., a retiree age 65 or older) who is eligible for some degree of help from Medicaid (which is a US federal/state government social welfare program). The retiree desires lifetime income and is concerned about health shocks that can lead to long term care, which is very expensive in the U.S. We assume that the retiree has a lump sum amount of G to buy a life annuity and pay for long term care. We explore the optimal design of a life annuity plus a long term care insurance rider to pay for long term care services and supports. To this end, we assume there are four stakeholders: (i) the retiree who wants to maximize her expected utility and minimize her out-of-pocket expenses, (ii) the long term care provider who wants to maximize its profits, (iii) Medicaid that wants to minimize its payments for long term care, and (iv) the insurer who wants to charge an actuarially sound premium. We develop the optimal design by constructing two sets of two dimensional multi-objective optimization problems. From the first set of optimization problems, the long term care provider and Medicaid provide the retiree with a set of policies along their Pareto frontier. Under the second set of optimization problems, the retiree creates her own Pareto frontier to make her optimal choice. An illustrative example is provided.

CI845 Room Beveridge Hall ADVANCES IN FORECASTING Chair: Michael Owyang

C0169: **Scenario forecasting with the conditional probit**

Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States

Elements from the literature on forecasting binary variables are merged with the literature on the construction and implementation of conditional forecasts. We do so by building on the Qual-VAR model. This model is a standard VAR but augmented with a continuous latent variable that, in turn, can be used to make probablitistic forecasts of a binary outcome as one would with a probit. The joint VAR-probit structure allows us to form counterfactual forecasts of the latent variable which can then be used to form probabilistic forecasts of the binary variable. We apply the model to forecasting recessions in real time and investigate the role of counterfactual monetary policy interventions on the likelihood of recessions.

C0179: **Measuring the effects of expectations shocks**

Presenter: **Ana Galvao**, University of Warwick, United Kingdom

It is shown that expectation shocks revisions in expectations unrelated to economic fundamentals have positive significant effects on US economic

activity. To measure the expectation shocks, we estimate a mixed-frequency VAR model that allows economic conditions in the current quarter to affect current-quarter GDP expectations. The model is estimated with real-time data so expectations shocks do not suffer a look-forward bias by incorporating future data revisions. Dynamic responses are estimated with the aid of a standard VAR. Expectations shocks explain 10% of the two-year variation of output, investment, consumption and hours. We find that expectations shocks are correlated with alternative belief-based shocks, but nevertheless have significant positive short-run effects on investment and hours even when the effects of the other shocks are controlled for.

C0182: **Tests of conditional predictive ability: Some simulation evidence**
Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States

Simple examples and associated simulations are used to investigate the size and power properties of tests of predictive ability. While we find that the tests can be accurately sized and powerful in large enough samples we identify details associated with the tests that are not otherwise apparent from the original text. In order of importance these include (i) the proposed test of equal finite-sample unconditional predictive ability is not asymptotically valid under the fixed scheme, (ii) for the same test, but when the rolling scheme is used, very large bandwidths are sometimes required when estimating long-run variances, and (iii) when conducting the proposed test of equal finite-sample conditional predictive ability, conditional heteroskedasticity is likely present when lagged loss differentials are used as instruments.

CO606 Room MAL B02 ECONOMETRIC STUDIES OF COMMODITY PRICES AND FUTURES Chair: Martin Stefan

C1079: **Global uncertainty and commodity futures dynamics**
Presenter: **Alexander Puetz**, Westfalische Wilhelms-Universitat Munster, Germany
Co-authors: Pierre Siklos

The possible effect of global uncertainty on the dynamics of daily futures prices of 30 commodities is investigated over the period from 1994 to 2019. It has been previously documented that high global uncertainty leads to a decline in real activity, while real activity is frequently associated with commodity price movements. The analysis is based on a Structural Dynamic Factor Model with a block structure to decompose each price series into common global components, sector specific block components and an idiosyncratic component. To measure global uncertainty, we employ several recently developed proxies.

C1588: **The influence of Brazilian exports on price transmission processes in the coffee sector: A Markov-switching approach**
Presenter: **Teresa Vollmer**, University of Goettingen, Germany
Co-authors: Stephan von Cramon-Taubadel

Most of the analysis of agricultural commodity market integration is solely based on price information. However, adding trade data can improve the understanding of interactions between interrelated markets. We link the analysis of price transmission processes between spot and futures markets with trade information to study the influence of Brazilian coffee exports on global price interdependencies. Using a Markov-switching vector error correction model (MSVECM) we allow for structural changes over time. The results reveal two regimes. One regime is characterized by periods of sideways or downward trending coffee prices with low price volatility, and the other one by phases of price spikes and high price volatility. Price information is transmitted through both the spot and the futures prices and the speed of the price transmission process is significantly affected by the total daily volume and value of Brazilian coffee exports.

C1733: **Testing for bubbles in commodity spot and futures using a co-explosive autoregression**
Presenter: **Neil Kellard**, University of Essex, United Kingdom
Co-authors: Sam Astill, Ioannis Korkos

Virtually all studies of asset price bubbles are carried out in a univariate framework. By contrast, we examine questions of bubble identification and consequently, market efficiency using a bivariate approach. Firstly, we apply a co-explosive vector autoregression to model whether the WTI crude oil price run up of 2007-2008 can be attributed to the existence of a bubble. We find evidence that there is (i) an explosive root in the system and (ii) that oil spot and futures prices at various maturities, are cointegrated over that period. Secondly, as an alternative approach, we apply recent univariate bubble tests to the difference between futures and spot prices. We finish with an evaluation regarding the most appropriate approach to bubble identification in commodity markets.

C1767: **Flight to quality: Gold mining shares versus gold bullion**
Presenter: **Philipp Prange**, Zeppelin University, Germany
Co-authors: Dirk Baur, Karsten Schweikert

Using the co-movement of gold mining shares with the price of gold, the strength of flight to quality and the severity of financial shocks are assessed by distinguishing between flight to physical gold and flight to gold mining companies. The analysis of a global sample of gold mining companies reveals that flights to quality are very different across financial shocks with the bankruptcy of Lehman Brothers and the Brexit vote being the most extreme at opposite ends of the spectrum. We also find evidence that a flight from gold mining shares leads to a stronger price reaction and thus safe haven effect of gold bullion. The findings demonstrate that gold mining companies can enrich our understanding of the flight to quality phenomenon.

C1884: **Who pays the piper calls the tune: Networks and transaction costs in commodity markets**
Presenter: **Christoph Sulewski**, WWU Munster, Germany
Co-authors: Pierre Siklos, Alexander Puetz

A new dataset of weekly wheat prices during the 1898 - 1914 is generated. Using variance decompositions from vector autoregressive models, a network of 9 wheat markets during the sample period is constructed and information spillovers between these markets are analyzed. Our results indicate that transaction costs are a significant determinant of the relative importance of market places in the continental European wheat trade.

CO412 Room MAL B04 MODELS OF DEPENDENCE, HEAVY TAILS AND FINANCIAL NETWORKS Chair: Rustam Ibragimov

C0342: **Oil prices and U.S. stock market dependence: A mixed-frequency data sampling copula approach**
Presenter: **Ruijun Bu**, University of Liverpool, United Kingdom

The relationship between oil prices and stocks is an important issue for portfolio selection and risk management. Understanding the economic factors affecting the interaction between oil prices and stocks allows investors to improve their portfolio performance. A mixed frequency data sampling copula model with explanatory variables (Copula-MIDAS-X) is proposed that incorporates low frequency explanatory variables into a high frequency dynamic copula model. The new model enables us to investigate the impacts of economic factors on the relationship between oil and stock returns. In an application to Brent oil prices and S&P 500 indices, we find that the dependence of oil and stock markets is influenced by aggregate demand and stock specific negative news. The impact of aggregate demand lasts for two years, while the impact of stock specific bad news lasts for one quarter. The implication for market regulators and investors is that changes in aggregate demand have influential and long-lasting effects on both oil prices and stock markets. Besides, investors who re-balance portfolios daily or weekly should use the information on both monthly economic indicators and daily returns in portfolio management.

C0370: **New measures of volatility clustering, nonlinear dependence and market (non-)efficiency**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and SPBU, Russia

Co-authors: Rasmus Soendergaard Pedersen, Rustam Ibragimov

Many key variables in finance, economics and risk management exhibit nonlinear dependence, heterogeneity and heavy-tailedness of some usually largely unknown type. Recent works in the literature have shown that heavy-tailedness the property of financial and economic markets is of key importance for robustness of many key models and standard inference approaches. The presence of non-linear dependence and heavy-tailedness may problematic the analysis of (non-)efficiency, volatility clustering and predictive regressions in economic and financial markets using traditional approaches based on ACFs of squared returns and asymptotic methods. Several new approaches are presented in order to deal with the above problems. The approaches are based on conservativeness properties of t-statistics and several new results on applicability of t-statistic based robust inference methods in the settings considered. In the approaches, estimates of parameters of interest (e.g., measures of nonlinear dependence based on sample autocorrelations of powers of the returns' absolute values) are computed for groups of data and the inference is based on t-statistics in resulting group estimates. Numerical results and empirical applications confirm advantages of the new approaches over existing ones and their wide applicability in the study of market (non-)efficiency, volatility clustering, nonlinear dependence, and other areas.

C0586: **A simple non-parametric goodness-of-fit test for elliptical copulas**

Presenter: **Miriam Jaser**, Technical University of Munich, Germany

Co-authors: Aleksey Min

A simple non-parametric goodness-of-fit test for elliptical copulas of any dimension is presented. It is based on the equality of Kendall's tau and Blomqvist's beta for all bivariate margins of an elliptical copula. First, the asymptotic distribution of the test statistic is derived under the assumption of known marginal distributions. Nominal level and power of the proposed test are investigated in a Monte Carlo study. In case of unknown marginal distributions, we estimate the margins non-parametrically and derive the limiting distribution using empirical copula processes. The limiting Gaussian field depends on the unknown copula. Therefore, we apply bootstrap procedures to conduct the goodness-of-fit test. Finally, an empirical application illustrates our goodness-of-fit test at work.

C0784: **Contagion in complex financial networks**

Presenter: **Kumushoy Abduraimova**, Imperial College London, United Kingdom

With economic integration global financial systems have been becoming more interconnected. Liberalisation of capital accounts, improved access to international capital markets, potentially better risk-sharing and many more are among the benefits of integration that the world has seen. The dark side is contagion though. We introduce a network-based contagion centrality measure that captures non-linear dependencies in extreme events. We apply it to analyse international stock markets contagion. The first observation is that contagion level has increased for all countries, advanced and emerging, during the Financial Crisis 2008. The contagion risk declines for all markets post-crisis, however remains above its pre-crisis level. The second finding is that advanced economies are more central in the global contagion network than the emerging ones. Advanced markets are more connected among each other and with the emerging markets, while the emerging ones do not show strong connection with the rest of emerging world. This resembles a so-called 'core-periphery' structure. Finally, the network effect captured by contagion centrality could potentially explain the tail risk of individual countries. More contagion-central countries have lower tail risk. They might be not very prone to tail risk, however, could have significant impact on the whole network conditional on the shock having occurred.

C1263: **Characterization of the tail behavior of a class of BEKK processes**

Presenter: **Rasmus Soendergaard Pedersen**, University of Copenhagen, Denmark

New, mild conditions are provided for strict stationarity and ergodicity of a class of multivariate BEKK processes. By exploiting that the processes can be represented as multivariate stochastic recurrence equations, we characterize the tail behavior of the associated stationary laws. Specifically, we show that each component of the BEKK processes is regularly varying with some tail index. In general, the tail index differs along the components, which contrasts most of the existing literature on the tail behavior of multivariate GARCH processes.

CO478 Room MAL B18 COMMODITIES FINANCE Chair: Julien Chevallier

C0319: **A dynamic conditional regime-switching GARCH CAPM for energy and financial markets**

Presenter: **Christian Urom**, Universite' Paris VIII Vincennes Saint Denis, France

Co-authors: Julien Chevallier

A methodology is developed for estimating a time-varying conditional version of the CAPM with regime changes in conditional variance dynamics. The goal is related to documenting the power of the beta when it is estimated dynamically. The conditional regime-switching GARCH CAPM, with time-varying betas explaining both bull and bear markets, outperforms the unconditional(static) CAPM. Among stocks, there are significant time variations in betas across our models and regimes. This empirical feature is even more pronounced in the USA, the UK, Germany, France, China, and Malaysia. Among energy and other commodities, we find similar variations in the market price of risk. The direction of the relation with market returns for Crude Oil, Gold, Copper, Tin, Rubber, Aluminum, and Platinum is the same across our nested models. This result also holds for aggregate markets indices. Secondly, we provide a ranking by mean filtered volatility series where Natural Gas stands out at a high level. Average pricing errors are inferior in the case of the conditional model, and for Crude Oil. Lastly, we demonstrate that the regime switching model delivers better estimates of one-day-ahead Value-at-Risk than its non-switching counterpart. Taken together, our results help shed light on the supremacy of the market factor alone associated with time variation in risk premia across energy and financial markets.

C0460: **A panel study of the effect of commodity price levels and volatility on the real exchange rates**

Presenter: **Nicola Rubino**, University of Barcelona, Spain

The aim is to study the impact of leading commodity prices long run term trends and their volatility on the real exchange rate short term convergence in an error correction background in a panel of developed and developing countries. Through the Mean Group DOLS estimator and the Logistic Smooth Transition Regression, we show that introducing a commodity price index in a cointegrating relationship with the real effective exchange rate instead of a standard terms of trade variable radically changes the long run impact of price variation, implying commodity dependency other than losses or gains in external competitiveness. The estimates show that emerging countries, and among those energy exporting ones, are those that would be more aggressively conditioned by disequilibria. As different measures of volatility are taken into account to capture arbitrage opportunities and the alternating regimes of convergence of the exchange rate to its equilibrium, it is proven that the Commodity Points theory of Hecksher can be effectively generalized at the longitudinal level without the need of resorting to external sources of variation which do not appear to have found their place in economic theory yet.

C0470: **Realized correlations, betas and volatility spillover in the agricultural commodity market: What has changed**

Presenter: **Matteo Bonato**, IPAG Business School, Switzerland

New insights are provided on the changes in the dynamics of price correlations and spillover effects in the commodity market. Using US-traded futures price data at a 1-minute frequency over the 2002- 2017 period, we consider the interaction within soft and grain commodities and between these commodities and oil. We rely on a recently introduced volatility model - the realized Beta GARCH model. Our results reveal that soft commodities were segmented prior to 2008 and became correlated thereafter. The nature of the increase in correlation is only temporary. The correlations within grains - already significant and positive - increased only marginally, indicating that this group has been less affected by recent events. The correlation between oil and agricultural commodities, which reached its peak in 2008, has also reverted to pre-crisis level. Spillover effects between oil and commodities have become more prominent prior to the commodity price crash. However, this increase in volatility

transmission tends to precede the increase in correlations. Finally, the impact of these findings on the performance of hedging strategies and optimal portfolio weights is discussed. Our results are important for investors exposed to the commodity market as they show that while the diversification benefits of investing in this market have decreased, volatility transmission risk and hedging costs have increased.

C0548: **Investigating bubbles in commodity prices by market expectations and determinants of dynamic persistence**
Presenter: **Christoph Wegener**, Leuphana University Lueneburg, Germany
Co-authors: Robinson Kruse-Becher, Tam Nguyen-Huu

Speculative bubbles have received a great deal of attention in the recent years by academics, policy makers and investors. One main research question relates to the important issue whether fundamentals are able to explain temporary explosive behavior in these commodity prices. We consider a set of seven unrelated commodities, i.e. wheat, copper, silver, sugar, soybean, cotton, and cattle, over the period between 1992 and 2019 at a monthly frequency. In a first step, we provide clear empirical evidence for multiple explosive episodes and find that these periods are cross-correlated. In a second step, we rely on model averaging techniques to explain the time-varying persistence of commodity prices with a large set of nearly two hundred potential determinants. The advantage of this approach is the avoidance of problems relating to the pre-specification of proxies for the fundamentals. As a comparison, we use a recently proposed approach which is based on market expectations of future prices. These two approaches are quite distinct in their nature. In our empirical analysis, we tackle the question whether dynamic persistence and explosive periods are related to fundamental explanations and provide corresponding conclusions.

C1935: **Behavioral, fundamentals and market determinants of the correlation between asset classes**
Presenter: **Yannick Le Pen**, Universite Paris Dauphine, France
Co-authors: Benoit Sevi

The correlations between stock, bond and commodity returns are of utmost importance for asset allocation but are also informative about the flight-to-quality issue. We estimate these correlations for a sample of US indices with the DCC-MIDAS model. This model distinguished between the long-run and a short-run component in the dynamics of conditional correlation. We extend the initial DCC-MIDAS to include an asymmetry effect of negative shocks on conditional standard deviations and correlations. In the next step we apply the definition of impulse response function in nonlinear models. In particular, we compute correlation impulse response functions (CIRF) in periods of high and low volatility as well as in periods of market downturn and upturn.

CO206 Room MAL B20 LARGE PANEL MODELS: ESTIMATION AND INFERENCE Chair: Kunpeng Li

C0326: **Time-varying income elasticities of healthcare expenditure for the OECD and the Eurozone**
Presenter: **Isabel Casas**, University of Southern Denmark, Denmark
Co-authors: Jiti Gao, Shangyu Xie, Bin Peng

Income elasticity dynamics of health expenditure is considered for the OECD and the Eurozone over the period 1995-2014. A new non-linear cointegration model is studied which has fixed effects, controlling for cross-section dependence and unobserved cross section and time heterogeneity. Most importantly, its coefficients can vary over time and its variables can be non-stationary. The resulting asymptotic theory is fundamentally different, with a faster rate of convergence, from similar kernel smoothing methodologies.

C0855: **Panel threshold models with interactive fixed effects**
Presenter: **Kunpeng Li**, Capital University of Economics and Business, China
Co-authors: Ke Miao, Liangjun Su

Estimation and inference in a panel threshold model is studied in the presence of interactive fixed effects. We study the asymptotic properties of the least squares estimators of the regression parameters in the model in the shrinking-threshold-effect framework. We find that under some conditions the threshold parameter can be estimated at a convergence rate faster than the usual parameter rate so that its estimation has asymptotically negligible role on the estimation of the slope coefficients in the model. The inference on the threshold parameter can be conducted based on a likelihood ratio test statistic as in the cross-sectional or time series setup. We also propose a test for the presence of threshold effect. Monte Carlo simulations suggest that our estimators and test statistics perform well in finite samples. We apply our method to study the effect of financial development on economic growth and find that there is indeed a turning point in the effect for all three measures of financial development when the cross-sectional dependence is properly accounted for.

C0893: **Functional-coefficient panel data models with cross-sectional dependence with an application to asset pricing**
Presenter: **Qiuhua Xu**, Southwestern University of Finance and Economics, China
Co-authors: Cai Zongwu, Ying Fang

A functional-coefficient panel data model with cross-sectional dependence is proposed motivated by re-examining the empirical performance of conditional capital asset pricing model. In order to characterize the time-varying property of assets betas and alpha, the proposed model allows the betas to be unknown functions of some macroeconomic and financial instruments. Moreover, a common factor structure is introduced to characterize cross-sectional dependence which is an attractive feature under a panel data regression setting as different assets or portfolios may be affected by same unobserved shocks. Compared to the extant studies, such as the classic Fama-MacBeth two-step procedure, our model can achieve substantial efficiency gains for inference by adopting a one-step procedure using the entire sample rather than a single cross-sectional regression at each time point. We propose a local linear common correlated effects estimator for estimating time-varying betas by pooling the data. The consistency and asymptotic normality of the proposed estimators are established. More importantly, an L2-norm statistic is constructed for testing the constancy of conditional betas and the significance of pricing errors. We show that the new test statistic has a limiting standard normal distribution under the null hypothesis.

C1751: **Large panel time series forecasting using functional model**
Presenter: **Mohammad Reza Yeganegi**, Islamic Azad University, Central Tehran Branch, Iran

Analysing Large panel data (panels with large number of corss-section units, N, and time periods, T) arises certain issues. Main challenges in large panel data analysis are Heterogeneity (i.e. fitting models with different parameters to each unit), Dynamics (i.e. using more complicated dynamic model since large amount of data is available for each unit, during time), Crosse-Section Dependency (i.e. modeling dependency between large number of units) and High Dimensionality (i.e. analysing the panel when N is considerably larger than T). There are different approaches to address these issues in large panel data, the proposed models usually does not address all the issues in the same time. For instance, whilst multivariate time series models and multivariate filters (e.g. State-Space models and Kalman filter) are powerful tools to address first three issues, they are not originally designed for high dimensional data. The potentials of functional time series models, as an approach to address above issues in forecasting large panel data, are investigated.

C1997: **Analyzing the social network with misspecification via double regularized GMM**
Presenter: **Chen Huang**, Aarhus University, Denmark
Co-authors: Weining Wang, Victor Chernozhukov

Social network analysis has gained significant attention recently. The identification, estimation and inference issues are intrinsically important in understanding the underlying network structure. We try to uncover the network effect with a predetermined adjacency matrix, and in addition we allow a flexible network specification by incorporating an unobserved network structure. In particular, the unobserved network structure can

be regarded as latent or misclassified network linkages. To achieve high quality estimator for parameters in both components, we propose to estimate via a double regularized high-dimensional GMM framework. We allow explicitly a factor structure in the instrument variables to address the ill-posedness of the inverse matrix involved. Moreover, this framework also facilitates us to conduct the inference. The theory of consistency and asymptotic normality is provided with accounting for general spatial and temporal dependency of the underlying data generating processes. Simulations demonstrate good performance of our proposed estimation and inference procedure.

CO378 Room MAL B35 NON-STANDARD ANALYSIS OF NON-LINEAR TIME SERIES Chair: Anurag Banerjee

C0209: **A novel approach to predictive accuracy testing in nested environments**
Presenter: **Jean-Yves Pitarakis**, University of Southampton, United Kingdom

A new approach is introduced to compare the predictive accuracy of two nested models that bypasses the difficulties caused by the degeneracy of the asymptotic variance of loss differentials used in the construction of commonly used predictive comparison statistics in the literature. The approach continues to rely on the out of sample MSE loss differentials between the two competing models, leads to Gaussian asymptotics and is shown to remain valid under flexible assumptions that can accommodate heteroskedasticity and the presence of mixed predictors (e.g. stationary and local to unit root). Simulations indicate that our methods have good finite sample properties.

C0233: **Mixed causal-noncausal autoregressions: Bimodality issues in estimation and unit root testing**
Presenter: **Frederique Bec**, THEMA University of Cergy-Pontoise and CREST, France
Co-authors: Heino Bohn Nielsen, Sarra Saidi

The bimodality of the widely used Student-t likelihood function applied in modelling Mixed causal-noncausal AutoRegressions (MAR) is stressed. It is first shown that a local maximum is very often to be found in addition to the global Maximum Likelihood Estimator (MLE), and that standard estimation algorithms could end up in this local maximum. It is then shown that the issue becomes more salient as the causal root of the process approaches unity from below. The consequences are important as the local maximum estimated roots are typically interchanged, attributing the noncausal one to the causal component and vice-versa, which severely changes the interpretation of the results. The properties of unit root tests based on this Student-t MLE of the backward root are obviously affected as well. To circumvent this issues, an estimation strategy is proposed which i) increases noticeably the probability to end up in the global MLE and ii) retains the maximum relevant for the unit root test against a MAR stationary alternative. An application to Brent crude oil price illustrates the proposed approach.

C0568: **Bandwidth choice in functional cointegration**
Presenter: **Xing Wang**, Durham University Business School, United Kingdom
Co-authors: Anurag Banerjee

The choice of the bandwidth in nonparametric predictive regression is studied. Two methods are considered: minimization of the MSE of the predictions, and choice based on the estimation of the convergence rate of the data generating process (DGP). The bin size k in piecewise local linear regression has been previously chosen by information criteria, which is an arbitrary method, while a rule of thumb $k = \sqrt{ln(T)}/T^{3/4}$ to achieve better MSE has also been proposed. We consider the bin size to be $k = T^{\alpha}$. Thus, k can be chosen by estimating α in a similar method to the one that is used to estimate the tail index of a time series. The potential advantage is twofold. First, the method is quite general in the sense that it can be applied to linear/nonlinear, integrated/long memory case. Second, an estimated α provide a clear approach to develop tests on the stationarity of the residuals in functional cointegration.

C0780: **Time-series GMM estimation with missing observations: A comparison of two alternative models**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan

The focus is on the two-step, efficient generalized method of moments (GMM) estimation of over-identified moment condition models using time-series data in which some observations are missing. The missing data problem arises for both high- (e.g., weekends and holidays for daily data) and low-frequency series (e.g., durations of two world wars for long-term annual or quarterly data). We investigate two alternative moment-condition models that can accommodate missingness, namely, the amplitude modulated and equal spacing models. The former assigns zeros in the place of missing observations, whereas the latter simply ignores missing observations and treats the observed data as if they were equally spaced in chronological order. We explore both large- and finite-sample properties of their corresponding efficient GMM estimators when the inverse of a kernel-smoothed heteroskedasticity and autocorrelation consistent (HAC) estimator is employed as the optimal weighting matrix in the second step. Moreover, a bandwidth selection method for kernel HAC estimation is proposed from the viewpoint of estimation optimality.

C1668: **Multiple long-run equilibria through cointegration eyes**
Presenter: **Jesus Gonzalo Munoz**, Universidad Carlos III de Madrid, Spain
Co-authors: Jun Yi Peng

Cointegration has succeeded in capturing the unique long-run linear equilibrium. Specific non-linearities have been incorporated into cointegrated models but always assuming the existence of a single equilibrium. We explore the possibility of different long-run equilibria depending on the state of the world (i.e., good and bad times, optimism and pessimism, frictional coordination) in a threshold framework. Starting from the present-value model (PVM) with different discount factors and depending on the state of the economy, we show that this type of PVM implies threshold cointegrated with different long-run equilibria. We present the estimation and inference theory. Two applications where the variables are not linearly cointegrated but threshold cointegrated are presented.

CO685 Room MAL B36 STATISTICAL LEARNING IN MACROECONOMICS AND FINANCE Chair: Henri Nyberg

C0484: **Boosting non-linear predictability of macroeconomic time series**
Presenter: **Timo Virtanen**, University of Turku, Finland
Co-authors: Heikki Kauppi

The boosting estimation method is applied to investigate to what extent and at what horizons macroeconomic time series have nonlinear predictability coming from their own history. The results indicate that the U.S. macroeconomic time series have more exploitable nonlinear predictability than previous studies have found. On average, the most favorable out-of-sample performance is obtained by a two-stage procedure, where a conventional linear prediction model is fine-tuned by the boosting technique.

C0511: **Directional predictability of daily stock returns**
Presenter: **Janis Becker**, Leibniz University Hanover, Germany
Co-authors: Christian Leschinski

In contrast to the monthly or quarterly case, daily stock returns are generally regarded as unpredictable. While this may be true for the level of daily returns, we focus on the signs of these returns. Using the logistic regression model, we show that meaningful directional forecasts can be generated. The analysis is carried out using pseudo out-of-sample forecasts for a data set consisting of all stocks in the Dow Jones Industrial Average from 2004-2017. Relevant predictor variables are chosen beforehand - in a separate model selection window from 1996-2003. This model selection procedure is carried out using a cross-validation procedure with forward chaining that is applicable in a time series context. Since the forecast period and the model selection period are strictly separated, the procedure mimics the Situation a forecaster would face in real time. Applying common statistical tests, our forecasts are shown to be statistically significant. Therefore, we draw the conclusion that the sign of daily stock

returns is (to some extent) predictable. Moreover, trading strategies based on these forecasts suggest the possibility to outperform the market index in terms of return and Sharpe ratio.

C0558: **How stock returns predictability using a simple neural network changes over time**
Presenter: **Adam Chudziak**, Szkola Glowna Handlowa w Warszawie, Poland

Scientific predictions in financial markets are commonly based on theoretical finance foundation. Nonetheless, many techniques used by practitioners do not come from theoretical considerations. Successful trading strategies based on Artificial Neural Networks (ANN) have been reported and are used by leading hedge funds. Although many ANN methods still used today, such as the multi-layer perceptron, have origins in 1950s and 1960s, the interest in them was rather small for almost half a century. Recently, more results on stock predictability using the ANN has been published, many of them use only past market behavior as predictor variables. However, they are usually constrained to specific time periods. The changes in the US stock prices monthly returns predictability are studied by using Artificial Neural Networks since the 1970s to 2015. We show that over this time there were periods when using past market data, even a simple feedforward Artificial Neural Network has some predictive capabilities. The investment strategy based on the predictions is tested. The predictability and profitability of trading varies between securities, but generally tends to decrease in time, with noticeable drop in performance in the 21st century. The results indicate that the stock returns predictability using ANN is not stable over time and that changing characteristics of the market require adaptation of forecasting techniques.

C0618: **Forecasting multinomial stock returns using machine learning methods**
Presenter: **Lauri Nevasalmi**, University of Turku, Finland

The daily returns of the S&P 500 stock market index are predicted using a variety of different machine learning methods. We propose a new multinomial classification approach to forecast stock returns. The multinomial approach can isolate the noisy fluctuation around zero and allows us to focus on predicting the more informative large absolute returns. Our in-sample and out-of-sample forecasting results indicate significant return predictability from a statistical point of view. Moreover, all the considered machine learning methods outperform the benchmark buy-and-hold strategy in a real-life trading simulation. The gradient boosting machine is the top-performer in terms of both the statistical and economic evaluation criteria.

C0897: **Semiparametric selection and optimal weighting of leading economic indicators**
Presenter: **Henri Nyberg**, University of Turku, Finland
Co-authors: Heikki Kauppi

A semiparametric econometric procedure is developed for getting maximal signalling predictive power out from a potentially large set of leading indicators to the state of the business cycle. The procedure is transparent, largely automated and meets demands from users working at different forecast horizons by selecting and weighting to obtain an optimal linear combination of leading indicators. The application with the U.S. data demonstrates the superiority of our procedure, providing a valuable complement to the existing methods and the benchmark composite index of leading indicators.

CO236 Room G4 NEW EMPIRICAL APPROACHES TO LONG RUN GROWTH Chair: Peter Pedroni

C0775: **Local population diversity**
Presenter: **Marc Klemp**, University of Copenhagen, Denmark
Co-authors: Oded Galor

Recent research has established that population diversity has had a profound effect on economic productivity over the course of human history. However, existing research is based on aggregate national level and ethnic level data and may confound the effects of unobserved national or ethnic characteristics with those of population diversity. Furthermore, the effect of within-ethnicity local diversity on long-run local development is unexplored. Accounting for unobserved national and ethnic characteristics, the effect of population diversity on long-run local development within a single nation and ethnicity is estimated. Focusing on the Canadian province Quebec, a novel measure of population diversity is developed which is based on the genetic distances between all pairs of individuals in the population as derived from comprehensive data on the genealogy of the entire historical population. Furthermore, the exogenous variation in local population diversity caused by the pattern of settlement of New France in the 17th century is exploited, and the effects of population diversity on the level of socioeconomic development two centuries later are estimated. Moreover, the persistence of local diversity, as well as the interaction effects of household-level and local-level diversity on individual socioeconomic prosperity, are exploited.

C0872: **Poverty, GDP per capita and health in developing countries: Panel cointegration results with selection**
Presenter: **Christophe Muller**, Aix-Marseille School of Economics, France
Co-authors: Cyrine Hannafi

Cointegration relationships between poverty, economic growth and health are estimated by using country-level data from developing countries, using Depth of Food Deficit, GDP per capita and Infant Mortality indicators as proxies. As opposed to the current literature, we account for non-stationarity issues, distinguish between short and long term interactions, and select countries for which all the indicators are non-stationary. The determinants for selection are explored with Probit models that show that, although the occurrence of stationarity is dominated by randomness, some shock variables (on food availability, measles and natural disasters) retain some prediction power of stationarity. We conduct a panel data cointegration analysis over the selected subsample. The VECM results are compared with those without sample selection and with those obtained for time series analyses for each country, and impulse response functions are examined. The cointegration results are found to vary with the estimation approach, and affected by sample selection. For a limited proportion of countries, the results exhibit strong serial correlations of poverty and infant mortality, and a substantial impact of the error correction term on the dynamics of all the studied variables. Finally, these statistical results are confronted with an economic theory model that provides an interpretation grid.

C1063: **The long-run dynamics of income and inequality**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Enrico Maria Cervellati, Gerrit Meyerheim, Uwe Sunde

The question of whether income inequality increases or decreases as a country develops, and what factors determine the shape of this relationship have been the matter of an intense debate over the past half century or more. We contribute to this debate in several ways. We propose a simple and coherent theoretical framework of long-run development that allow us to investigate the mechanisms behind the long-run dynamics of income and the distribution of income. The model generates dynamics of inequality in various dimensions over the long run and along the entire income distribution, both within and across countries as the result of technological change and its interplay with demographic development. The analytical framework reconciles seemingly incompatible and contradictory empirical results such as the Kuznets curve and the recent increase in inequality due to changes at the top tail within a single theoretical framework. By combining the analytical insights with a new econometric methodology for robustly estimating nonlinear relationships, we demonstrate a new approach to studying long run dynamics in the context of the model and investigate the role that institutions and demographics have played in shaping the income income-inequality nexus.

C1010: **The long-run dynamics of capital-skill complementarity**
Presenter: **Uwe Sunde**, LMU Munich, Germany
Co-authors: Gerrit Meyerheim, Peter Pedroni

The question about the relative complementarity of capital and skills is addressed. While there is an emerging consensus that physical capital and skills are relative complements in advanced economies, economic historians have argued that, at least during the early stages of the industrial revolution, capital and unskilled labor have been relatively more complementary than capital and skilled labor. Relatively little is known empirically about the complementarity of capital and skills in developing countries and over the long-run. One reason is the lack of an appropriate empirical methodology with which the relationship can be estimated robustly while being sufficiently flexible to be able to account for changes in this relationship and identify potential determinants. We investigate the relationship from the perspective of a model of long-run growth and employ a novel econometric methodology that can be used to estimate the relationship between income and capital inputs depending on the available skill composition of the labor force. The methodology also allows us to isolate changes in this relationship for different levels of capital and skills and to provide estimates of non-functional relationships between different variables.

C1110: **Is economic growth really over? A contribution to the secular stagnation debate**
Presenter: **Gerrit Meyerheim**, LMU Munich, Germany
Co-authors: Peter Pedroni, Uwe Sunde

The observation of a slow-down in trend growth in almost all developed economies over the past decades has sparked a debate about the reasons for this secular stagnation. A prominent interpretation has been to associate the stagnation with a decline in productivity growth and an excessive accumulation of capital over the same period. Several potential factors have been held responsible for these empirical patterns, including demographic factors related to population aging, education, inequality, globalization, environmental factors, and public finances, but the discussion of the empirical relevance of these arguments remains largely descriptive. The relationship between income and productivity are investigated from the perspective of a long-run growth framework that implies non-linear long run dynamics as economies experience a transition from stagnation to sustained growth. We pair this framework with a novel empirical methodology to estimate the non-linear relationships between income, productivity and capital stipulated by the model and explore the empirical relevance of the different arguments that have been emphasized in the literature.

CO472 Room Gordon ECOSTA JOURNAL PART A: ECONOMETRICS Chair: Alessandra Amendola

C1942: **Combining different frequencies in modeling non–negative processes: The MEM-MIDAS**
Presenter: **Giampiero Gallo**, NYU in Florence, Italy
Co-authors: Alessandra Amendola, Vincenzo Candila, Fabrizio Cipollini

In modeling financial time series, information of interest may be available at different frequency of observation. We extend the MIDAS-GARCH model to explicitly take in consideration that a multiplicative error model may be a more direct way to model the conditional expectation of a non-negative process observed daily and that a low frequency component in the data can be modeled exploiting some other information sampled at a different frequency, say monthly. The empirical application is presented on the realized volatility of the Dow Jones 30 components and the S&P500.

C0683: **Seasonal adjustment of high-frequency data**
Presenter: **Jan Jacobs**, University of Groningen, Netherlands
Co-authors: Barend Abeln

In the last decade large data sets have become available, both in terms of the number of time series and with higher frequencies (daily and even higher). All series may suffer from seasonality, which hide other important fluctuations. Therefore time series are typically seasonally adjusted. Recently, CAMPLET, a new seasonal adjustment method which does not produce revisions when new observation become available, has been presented. The aim is to show the attractiveness of CAMPLET for high frequency time series. We illustrate seasonal adjustment in CAMPLET for the series of daily consumption of electricity in France.

C1655: **The portfolio regression approach to mean-variance analysis via the elastic net**
Presenter: **Red Laviste**, University of Basel, Switzerland
Co-authors: Dietmar Maringer

The portfolio regression approach encompasses models that estimate mean-variance efficient portfolio weights using multiple linear regression. Being the regression coefficients, estimates of the portfolio weights obtained from the regression approach are equivalent to the more famous portfolio optimization approach, but can be assessed with standard errors and other well-known statistical measures and tests. Since the squared residual loss in least squares regression has a tendency to overfit, regularized regression models such as ridge and lasso are often employed to induce stability and sparsity in portfolio weight estimates. We unify existing portfolio regression models into an elastic net framework that nests the global minimum variance, tangency and frontier portfolios as special cases, and study the optimal calibration of the regularization factor λ.

C1725: **Sparse index tracking via the sorted ℓ_1 - norm**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Philipp Johannes Kremer, Damian Brzyski, Malgorzata Bogdan

Index tracking and hedge fund replication aims at replicating or cloning the risk-return properties of a given benchmark, by either using only a subset of its original constituents or by a set of risk factors. We propose a new statistical model for index tracking and hedge fund replication, that relies on the convex *Sorted ℓ_1 Penalized Estimator* (SLOPE). SLOPE is capable not only to provide sparse clones but also to automatically group assets sharing similar statistical properties with respect to the benchmark, and thereby allowing to develop further investment strategies.

C1556: **Dynamic quantile function models**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Wilson Chen, Gareth Peters, Scott Sisson

A novel way of thinking about the modelling of the time-varying distributions of financial asset returns is presented. Borrowing ideas from symbolic data analysis, data representations beyond scalars and vectors are considered. Specifically, a quantile function is considered as an observation, and a new class of dynamic models for quantile-function-valued (QF-valued) time series is developed. In order to make statistical inferences and account for parameter uncertainty, a method whereby a likelihood function can be constructed for QF-valued data is proposed, and an adaptive MCMC sampling algorithm for simulating from the posterior distribution is developed. Compared to modelling realized measures, modelling the entire quantile function of intra-daily returns allows one to gain more insight into the dynamic structure of price movements. Via simulations, we show that the proposed MCMC algorithm is effective in recovering the posterior distribution. In the empirical study, the new model is applied to analyze one-minute returns for major international stock indices. Through quantile scaling, we further demonstrate the usefulness of our method by forecasting one-step-ahead the Value-at-Risk of daily returns.

CO785 Room Montague UNDERSTANDING THE CROSS SECTION OF STOCK RETURNS Chair: Julien Penasse

C0260: **Estimating the anomaly baserate**
Presenter: **Andreas Neuhierl**, University of Notre Dame, United States
Co-authors: Alex Chinco, Michael Weber

The academic literature contains literally hundreds of variables that seem to predict the cross-section of expected returns. This so-called anomaly zoo has caused many to question whether researchers are using the right tests for statistical significance. But, here is the thing: even if a researcher

is using the right tests, he will still be drawing the wrong conclusions from his analysis if he is starting out with the wrong priors, i.e., if he is starting out with incorrect beliefs about the ex ante probability of discovering a tradable anomaly prior to seeing any test results. So, what are the right priors to start out with? What is the correct anomaly baserate? We propose a new statistical approach to answer this question. The key insight is that, under certain conditions, there is a one-to-one mapping between the ex ante probability of discovering a tradable anomaly and the best-fit tuning parameter in a penalized regression. When we apply our new statistical approach to the cross-section of monthly returns, we find that the anomaly baserate has fluctuated substantially since the start of our sample in May 1973. The ex ante probability of discovering a tradable anomaly was much higher in 2003 than in 1990. As a proof of concept, we construct a trading strategy that invests in previously discovered predictors and show that adjusting this strategy to account for the prevailing anomaly baserate boosts its performance.

C0255: **Understanding alpha decay**
Presenter: **Julien Penasse**, University of Luxembourg, Luxembourg

The relationship between realized anomaly returns and expected risk-adusted anomaly returns (alpha) is clarified. When the alpha of an anomaly decays, for example after the anomaly has been discovered and traded on, the portfolio's market value increases leading to outsized positive realized returns. That is, the average of (recent) past realized returns leads to an overestimation of expected returns (alpha) going forward. We show that ignoring this negative correlation between expected returns and realized returns can meaningfully affect statistical inference in the anomalies literature and we provide a simple formula that corrects for this effect.

C0264: **A critique of momentum anomalies**
Presenter: **Thiago de Oliveira Souza**, University of Southern Denmark, Denmark

Theoretical, empirical, and simulated evidence is offered showing that momentum regularities in asset prices are not anomalies. Within a general, frictionless, rational expectations, risk-based asset pricing framework, riskier assets tend to be in the loser portfolios after (large) increases in the price of risk. Hence, the risk of momentum portfolios usually decreases with the prevailing price of risk, and their risk premiums are approximately negative quadratic functions of the price of risk (and the market premium) theoretically truncated at zero. The best linear (CAPM) function describing this relation unconditionally has exactly the negative slope and positive intercept documented empirically.

C0649: **Forest through the trees: Building cross-sections of stock returns**
Presenter: **Svetlana Bryzgalova**, London Business School, United Kingdom
Co-authors: Markus Pelger, Jason Zhu

Sorting-based strategy of building portfolios has been a default empirical approach in asset pricing for creating both test assets and factor-mimicking returns. One of the natural limitations of this technique, however, is its inability to adequately reflect the information contained in more than 2 characteristics and their interaction. Yet recent advances in empirical asset pricing have repeatedly highlighted the importance of the latter. We propose to analyze the effect of a large number of characteristics on expected stock returns with the machine learning technique known as random forest. As an ensemble learning method for classification, the new approach is particularly well-suited for building composite cross-sections of portfolios that reflect the rich conditional information contained in a large number of characteristics simultaneously, and can be viewed as a natural generalization of the conventional sorting based strategies. We build decision trees for various sets of stock specific characteristics, and demonstrate that the new approach is able to create cross-sections that a) reflect the information in a joint conditional distribution of characteristics, b) are challenging to price based on the conventional models, even when pitted against the tradable factors based on the underlying characteristics, and c) imply investment strategies that achieve yearly out-of-sample annual Sharpe ratios above 2.

C1049: **Linear factor models and the estimation of expected returns**
Presenter: **Cisil Sarisoy**, Federal Reserve Board, United States
Co-authors: Peter de Goeij, Bas Werker

The focus is on analyzing the properties of expected return estimators on individual assets implied by the linear factor models of asset pricing, i.e., the product of beta and lambda. We show that using factor-model-based risk premium estimates leads to precision gains of up to 31% when compared to the historical averages. In the presence of omitted factors, adding an alpha to the model captures mispricing only in case of traded factors, otherwise the bias caused by misspecification can not be corrected. Finally, inference about expected returns, unlike inference on factor prices, does not suffer from a small-beta bias. The more precise factor-model-based estimates of expected returns translate into significant improvements in out-of-sample performance of optimal portfolios.

CO470 Room Woburn APPLIED INTERNATIONAL MACROECONOMICS Chair: Mariarosaria Comunale

C0312: **Shock dependence of exchange rate pass-through: A comparative analysis of BVARs and DSGEs**
Presenter: **Mariarosaria Comunale**, Bank of Lithuania, Lithuania

Results from Structural Bayesian VARs coming from several studies for the euro area are collected which apply the idea of a shock-dependent Exchange Rate Pass-Through, drawing a comparison across models and also with respect to available DSGEs. On impact the results are similar across Structural Bayesian VARs. It is, however, very hard to find a robust characterization across models and the modelling challenges increase when looking at individual countries. Hence, we provide a local projection exercise with common euro area shocks, identified in euro area-specific Structural Bayesian VARs, extrapolated and used as regressors. For common exchange rate shocks, the impact on consumer prices is the largest in some new member states, but there are a wide range of estimates across models. Generally euro area monetary policy plays a bigger role for consumer prices. The very low values in core consumer prices can be mostly attributed to the price of services. In BVARs especially, all shocks contribute relatively little to observed changes in the exchange rate and in HICP, pointing to a key role of the contribution of systematic factors, not captured by the historical shock decomposition.

C0344: **Shocking interest rate floors**
Presenter: **Daniel Kaufmann**, University of Neuchatel, Switzerland

The dynamic causal effects of interest rate floor shocks are identified exploiting regular auctions of Swiss central bank debt securities (SNB Bills). We first establish theoretically that central bank debt and interest-bearing reserves are equivalent when reserves are ample. Based on these insights, the empirical analysis identifies an interest rate floor shock in a dynamic event study of SNB Bill auctions. A restrictive interest rate floor shock causes an increase in the money market rate, a persistent appreciation of the Swiss franc, a decline in long-term interest rates, and a decline in stock prices. We then perform policy experiments under various identifying assumptions in which the central bank raises the interest rate floor from 0% to 0.25%. Such a policy change causes a 3-6% appreciation of the Swiss franc and a 5-20% decline in stock prices.

C0488: **On how firms adjust when trade stops: Labor market, industrial linkages, and macroeconomic effects**
Presenter: **Povilas Lastauskas**, CEFER Vilnius University, Lithuania
Co-authors: Aurelija Proskute, Alminas Zaldokas

The aim is to investigate how firms adjust to sudden and unanticipated negative trade shock coming from the sanctions completely banning exports to one of their major trade partners. We explore a unique event, Russia banning imports of agricultural and food products from the European Union in 2014, when due to political reasons, unrelated to trade, the exporters in a small open economy have lost access to a major export market. The negative trade shock that we explore allows us to quantify the firm adjustment margins at the micro level and their propagation into the macro economy via the input-output network structure and the implied linkages among the firms. Abstracting away from a number of other adjustment margins

or employment options, we instead assume empirically relevant ingredients: full- and part-time work, worker heterogeneity, lower employment costs. We find that firms have primarily adjusted by reducing their full-time employment by shifting to the part-time employment, both in terms of the number of people and the hours worked. However, the average wage per hour (or per employee) has not decreased (rather, the opposite), thus indicating that the remaining employees had the same (or higher) productivity. This suggests that trade policies have an immediate effect on labor market and that recent global trends reversing trade integration might have adverse consequences on labor markets in the export-dependent industries.

C1609: **Revisiting the manufacturing-led growth hypothesis: A quantile regression approach**
Presenter: **George Voucharas**, University of Macedonia, Greece
Co-authors: Theodore Panagiotidis

The export-Led growth hypothesis is revisited by employing a panel set of 120 countries covering annual data over the period 1980-2017. Total exports are disaggregated into primary and manufacturing exports, and control variables are taken into account. The assumption of symmetry is relaxed and a panel quantile regression framework that can quantify the effects of manufacturing exports on relatively poor and relatively rich countries is established. The entire conditional distribution of income is modelled, taking into consideration the unobserved heterogeneity and endogeneity concerns. Although the effect of total exports on growth does not vary across quantiles of income, poor and developing countries benefit more from manufacturing exports than the rich and developed ones.

C1590: **Impact of regional trade agreements on trade creation and trade diversion using the structural gravity model**
Presenter: **Barton Sy**, TAT SING International Logistics Corporation, Philippines
Co-authors: Stephen Jun Villejo, Rutcher Lacaza

The aim is to examine the impact of regional trade agreements such as the ASEAN Free Trade Agreement (AFTA) using the structural gravity model. The empirical analysis uses trade data of member countries of ASEAN, APEC, and EU from 1990 to 2014. Estimation of multilateral resistances used the fixed effects method. Because of zero trade flows and heteroscedasticity in the data, nonlinear methods are employed, namely the Helpman two-step selection process method and the Poisson Pseudo Maximum Likelihood approach. In addition to using exporter-time and importer-time fixed effects, pair fixed effects are also used to account for the endogeneity of the trade policy variables. A methodology to test for potential reverse causality, and non-linear and phasing-in effects of trade agreements and policies is implemented. Results show that the AFTA is a significant positive determinant for trade creation and does not cause trade diversion, though this is not the case for the other regional trade agreements. The results also confirm the absence of reverse causality. Lastly, the results show that the trade agreements have a non-linear phasing-in effect wherein the lag 3 effect has the highest magnitude.

CO244 Room Chancellor's Hall ENVIRONMENTAL ECONOMETRICS Chair: Simone Maxand

C0616: **Econometrics for climate modelling**
Presenter: **David Hendry**, Oxford, United Kingdom
Co-authors: Jennifer Castle

Climate time series are non-stationary from both stochastic trends, tackled by cointegration, and distributional shifts-caused by everything from volcanic eruptions to policy interventions-tackled by indicator saturation. The tools include [a] model selection retaining theory information while selecting over other candidate variables; [b] software that can handle more variables, N, than observations, T, by expanding and contracting multi-path block searches; and [c] saturation estimation, including in that candidate set indicators for a range of potential contaminations, including outliers (impulse-indicator saturation, IIS), location shifts (step-indicator saturation, SIS), both (IIS+SIS, super saturation), changes in model parameters (multiple-indicator saturation, MIS), and designed indicators (e.g.) for impacts of volcanic eruptions on temperatures (DIS). Despite $N > T$ (often several fold), the costs of selection are small relative to mis-specification problems that might otherwise occur. Areas of application include modelling UK CO2 emissions from 1860 onwards.

C1288: **Modelling emissions by saturation estimation**
Presenter: **Jonas Kai Kurle**, University of Oxford, United Kingdom

Super saturation (impulse indicator saturation and step indicator saturation) for modelling outliers and location shifts in statistical processes is highly relevant for environmental and climate econometrics. The properties of this approach are examined through extensive Monte Carlo simulations. They show that for tight significance levels, the detection rate of these breaks is generally high while the retention rate of irrelevant indicators is well-controlled. Furthermore, a newly designed indicator to model smooth location shifts is introduced, which is called policy transition indicator saturation (PTIS). Compared to step indicator saturation, the potency tends to be lower, which is partly due to larger uncertainty of detecting the correct break date. An application to UK CO2 emissions shows that PTIS may improve empirical modelling.

C1406: **Modelling long-run co-volatilities**
Presenter: **Susana Martins**, University of Oxford, United Kingdom
Co-authors: David Hendry

Long-run co-volatility between climate and financial variables is of increasing interest as the climate changes seem to be happening much faster than the IPCC models initially expected. There is a strong possibility that this will wrong foot markets. There have been studies showing that the volatility of some climate variables, such as the jet stream, has been changing greatly, but modelling co-volatilities is a new research area.

C1159: **A panel approach for causalities and the distribution between income inequality and carbon emissions**
Presenter: **Franziska Dorn**, University of Goettingen, Germany
Co-authors: Simone Maxand, Thomas Kneib

High levels of carbon emissions and rising income inequality are two of the greatest challenges in the global society. Recent literature has investigated this relationship by primarily establishing linear regressions. The aim is to add twofold to the debate of causal relationship and the interdependence between carbon dioxide emissions and income inequality. First, a panel vector autoregressive regression model is implemented to establish causal inference between carbon dioxide emissions and income inequality. This technique enables to model the dynamic bilateral relation between the two dimensions and accounts for varying time effects. Second, distributional copula models are used to analyze the conditional dependence between the two dimensions. By using generalized additive models for location, scale and shape, the variation of the covariates along the distribution of the outcome variable is analyzed. First results show that the dependence and causalities between the two dimensions vary by different macro-economic contexts of countries. A comparison of high-, middle-, and low- income countries indicates that the dependence structure changes with the level of income. However, the relationship is negative in all countries, which states that gains in one dimension come at the cost of the other.

C1022: **Economic growth, energy consumption and carbon emissions: Evidence from statistically identified panel SVARs**
Presenter: **Simone Maxand**, University of Helsinki, Finland

In the light of intensifying discussions on climate risks, the detection of global and local causal structures between economic and environmental variables becomes increasingly interesting. In this respect, the outcome is twofold. First, we advance panel VAR methodologies by statistical identification based on group independent component analysis techniques. On different cross-section levels this allows for the interpretation of either country- or region-specific structural shocks. In a second step, these panel SVAR techniques provide a powerful tool to derive new insights

on the interconnection of economic growth, energy consumption and carbon emissions at global and regional levels.

CC815 Room MAL 352 CONTRIBUTIONS IN TIME SERIES ECONOMETRICS Chair: Alessandra Luati

C1193: **Jointly modeling autoregressive conditional mean and variance of positive valued time series**
Presenter: **Hiroyuki Kawakatsu**, Dublin City University, Ireland

Observation driven models with conditional mean and variance dynamics are proposed for positive valued time series. The motivation is to relax the strong restriction on the higher order moment dynamics implied by the standard mutiplicative error model that is driven only by the conditional mean dynamics. The empirical fit of the proposed specifications is assessed with daily realized volatility series for a number of stock indices using both in-sample estimates and pseudo out-of-sample prediction densities.

C1583: **Dornbusch's overshooting revisited: A bounce-back threshold autoregression with ARCH effect**
Presenter: **Melika Ben Salem**, University Paris-Est Marne-la Vallee, France
Co-authors: Frederique Bec

Excessive exchange rate variation could be explained by the overshooting effect first identified in the seventies. Even though more sophisticated versions of that model have been proposed since then, this monetary description of the exchange rate behavior, whose influence remains important, has received little empirical support in a linear multivariate framework. More recently, in another branch of literature, bounce-back augmented nonlinear models have been found to be useful to describe transitory epochs of high growth rate GDP recovery following a recession and preceding a normal growth rate regime. The idea is to bring these two strands of research together to shed new light on the nonlinear modeling of nominal exchange rates. The contribution consists in evaluating the ability of a bounce-back effect augmented nonlinear threshold autoregression, extended to allow for an ARCH component, to capture this overshooting behavior. First results using monthly nominal GBP/USD exchange rate data since 1970 provide evidence that (i) the null hypothesis of no bounce-back effect is strongly rejected and (ii) the estimated values of parameters of the bounce-back function are in line with the overshooting hypothesis.

C1709: **Forecasting regional inflation using spatial correlation models**
Presenter: **Taisiia Gorshkova**, Russian Presidential Academy of National Economy and Public Administration, Russia

The purpose is to discuss the need to integrate the spatial relationship in regional data for forecasting inflation. There is a comparative analysis of the models that take into account only temporal correlation between the data and models that take into account temporal and spatial correlation. Three weighting matrices are used to consider spatial correlation in data and six different models with each weight matrix, such as individual models for each region, panels with fixed and random effects, spatial lag models and spatial error models. The individual forecasts based on the models were then combined with different weights into a single forecast. The weights were chosen on the basis of five methods, including discounting and shrinkage methods and method of principal component analysis.

C1695: **A robust approach to heteroskedasticity, serial correlation and slope heterogeneity for large linear panel data models**
Presenter: **Kazuhiko Hayakawa**, Hiroshima University, Japan
Co-authors: Guowei Cui, Shuichi Nagata, Takashi Yamagata

A robust approach is proposed against heteroskedasticity, error serial correlation and slope heterogeneity for large linear panel data models. First, we establish the asymptotic validity of the Wald test based on the widely used panel heteroskedasticity and autocorrelation consistent (HAC) variance estimator of the pooled estimator under random coefficient models. Then, we show that a similar result holds with the proposed bias-corrected estimator for models with unobserved interactive effects. Our new theoretical result justifies the use of the same slope estimator and the variance estimator, both for slope homogeneous and heterogeneous models. This robust approach can significantly reduce the model selection uncertainty for applied researchers. In addition, we propose a novel test for the correlation and dependence of the random coefficient with covariates. The test is of great importance, since the widely used estimators and/or its variance estimators can become inconsistent when the variation of coefficients depends on covariates, in general. The finite sample evidence supports the usefulness and reliability of our approach.

C1916: **Nowcasting monthly GDP with big data: A model averaging approach**
Presenter: **Alessandro Giovannelli**, University of Rome Tor Vergata, Italy
Co-authors: Tommaso Proietti

Gross domestic product (GDP) is the most comprehensive and authoritative measure of economic activity. The macroeconomic literature has focused on nowcasting and forecasting this measure at the monthly frequency, using related high frequency indicators. The issue of estimating monthly gross domestic product is addressed by using a large dimensional set of monthly indicators, by pooling the disaggregate estimates arising from simple and feasible bivariate models that consider one indicator at a time, in conjunction to GDP or a component of GDP. The weights used for the combination reflect the ability to nowcast the original quarterly GDP component. The base model handles mixed frequency data and ragged-edge data structure with any pattern of missingness. The methodology allows us to assess the contribution of the monthly indicators to the estimation of monthly GDP, thereby providing essential information on their relevance. This evaluation leads to several interesting discoveries.

CG389 Room MAL 351 CONTRIBUTIONS IN MONETARY POLICY Chair: Par Osterholm

C1754: **On the effect of monetary policy on bank behaviour: Evidence from bank credit standards**
Presenter: **Nektarios Michail**, Cyprus University of Technology, Cyprus
Co-authors: Demetris Koursaros

The purpose is to examine whether conventional monetary policy has an impact on bank credit standards through the manipulation of interest rates. Using three distinct methodologies, the results confirm that the policy rate appears to have the expected tightening impact on credit standards. However, this effect is not found to be large and, most importantly, it is likely to be outweighed by the presence of counteracting factors, the most notable of which is private consumption. Other macroeconomic factors such as the yield curve, inflation, investment, and housing prices also have an impact on bank credit standards, the size of which varies across specifications. As such, the empirical results suggest that while the interest rate can cool off banking behaviour, ceteris paribus, i.e. if no change in the economy takes place, this is not likely a realistic scenario given that many other changes in other macroeconomic factors also take place at the same time.

C1829: **The impact of heterogeneous unconventional monetary policy on tail risks**
Presenter: **Antoni Vaello Sebastia**, University of Balearic Islands, Spain
Co-authors: Irma Alonso, Pedro Serrano

The impact of unconventional monetary policies (UMPs) of four major central banks -US, Japan, Europe and UK- on market uncertainty is analyzed. We exploit the heterogeneity of different UMP actions to disentangle a differential impact of these measures on option-implied risk neutral densities. Using an event-study, the preliminary results show that the announcement of UMP generally reduces the option-implied probability of risky events across different horizons and thresholds for a given loss, suggesting that the risk-taking channel have worked in the four areas analysed. Most of the results seem to be driven by forward guidance and liquidity measures rather than asset purchases. Cross-border effects are also relevant, but they only affect larger horizons, which suggests the existence of a differential impact depending on whether there is an idiosyncratic or exogenous contribution of UMP. Finally, the dynamics of the UMP processes are captured by a SVAR using sign restrictions and differentiating between an unconventional monetary policy, demand, supply and uncertainty shock.

C1864: **Preferred habitat, policy, and the CIP puzzle**
Presenter: **Paul Wohlfarth**, Birkbeck University of London, United Kingdom

The purpose is to examine the impact of policy and market segmentation on the failure of covered interest parity, CIP, a crucial no-arbitrage condition in international finance. The framework integrates market segmentation on fixed income markets, using preferred habitat theory, into an augmented CIP condition with intermediation costs. It highlights the interplay between policy, risk, and structural factors and the role of arbitrage in an endogenous channel for intermediation frictions. We estimate policy transmission channels in EGARCH-in-Mean models of USD/EUR cross currency basis swap (CCBS) rates across maturities. Our findings provide evidence for direct and indirect policy transmission effects on FX swap markets, affecting means and variances. Analysing co-movement across the CCBS term-structure in a VECM framework provides evidence for time-varying market segmentation that appears to be linked to volatility. Our findings highlight the importance of risk- and policy factors in conjunction with structural intermediation factors for foreign exchange markets.

C1867: **Macroeconomic interactions with money in China**
Presenter: **Xiaohong Chen**, Birkbeck, University of London, United Kingdom

The aim is to estimate macroeconomic models of the Chinese economy to analyse the links between money and income, exports, inflation and interest rates. As China's economic reforms have undergone significant structural breaks after 1979 and 1992, 5-variable vector auto-regression, VAR estimated on two periods, 1980Q1-1992Q4 and 1993Q1-2018Q3, is used. The empirical evidence shows a long-run, cointegrating, money demand function in which the estimated long-run real income and real interest elasticity are respectively 1.51 (1.46) and -0.04 (0.01) over the period 1980Q1-1992Q4 (1993Q1-2018Q3). It also shows a long-run income equation in which the estimated long-run real exports and real interest elasticity are respectively 0.7 (0.68) and 0.01 (-0.03) over the period 1980Q1-1992Q4 (1993Q1-2018Q3). But the impact of real interest rate on long-run income is only significant from 1993.

C1819: **Less bang for the buck? Assessing the role of inflation uncertainty for U.S. monetary policy transmission**
Presenter: **Hannes Rohloff**, University Goettingen, Germany
Co-authors: Helmut Herwartz

The relationship between inflation uncertainty and monetary policy transmission in the U.S. economy is investigated. Monetary policy shocks are identified within the framework of nonlinear structural factor-augmented VARs which allow us to analyze several complementary hypotheses connecting inflation uncertainty with reduced monetary policy effectiveness. We find that the real effects of monetary policy shocks are markedly dampened conditional on high inflation uncertainty. This can be traced back to, inter alia, real-option and precautionary savings effects which distort the traditional interest rate channel. Moreover, policy transmission through the external finance premium and the term structure of interest rates appears strongly dependent on inflation uncertainty.

CG865 Room MAL 353 CONTRIBUTIONS IN CREDIT RISK Chair: Jonathan Crook

C0293: **Financial frictions and housing collateral constraints in a macro model with heuristics**
Presenter: **Corrado Macchiarelli**, Brunel University, United Kingdom
Co-authors: Paul De Grauwe

The role of household debt in the real activity has attracted considerable attention recently mostly in the light of the observed increases in property prices and the increase of household indebtedness prior to the 2008 bust in many countries. The relevant literature on housing points to a number of the mechanism being likely to trigger or amplify real estate cycles, including bubbles. We focus on the interaction between banks and real estate developments, in particular assessing the implications of changing property prices on consumption decisions. We build on a previously described framework to introduce a real estate sector, accounting in itself for an explicit balance sheet dimension for consumers. The model thus results in an economy where - on the demand side - a collateral constraint limits households ability to borrow against the value of the real estate, and - on the supply side - loan supply is constrained by bank capital. This allows studying the interactions of these two limits by drawing a stark distinction between the supply and demand for credit. While lending constraints are not a new feature of this framework, we take a step further and analyse the implications of lending constraints in a bounded rationality framework. Together with considering bounded rationality rules, the model features an endogenous mechanism for describing the probability distribution of housing bubbles.

C0751: **Multi-factor model for contingent convertible bonds**
Presenter: **Renata Latocha**, ODDO BHF Asset Management GmbH, Germany

The aim is to identify the cross sectional multi-factor structural model that provides insight into the drivers of prices of contingent convertible (CoCo) bonds. The OLS regression performance with cross-sectional data show the effectiveness of this method when dealing with CoCos market which is too immature to draw any far reaching conclusions on modeling and calibration from the market data. The OLS estimator minimizes the squared distance between the plane with multiple factor loadings in the multidimensional matrix and the regressand vector of CoCos prices. The identified model retains predictive power for the portfolio of 9705 global corporate bonds including CoCos. The values of estimated parameters of the bonds prices functions indicate that the risk-free interest rate, maturity of the bond, credit spread, coupon level, current economic sytuation and expected one, discount factor and distance-to-default factor influence in prices of CoCos significantly. The CoCos specific factors like trigger spread or conversion type have moderate affect the prices of CoCos. The results suggest that the constructed model is a pragmatic approach to analyze and explain CoCos prices.

C1649: **Credit rating downgrade risk on equity returns**
Presenter: **Periklis Brakatsoulas**, Charles University, Faculty of Social Sciences, Czech Republic

A four-factor model directed at capturing the size, value, and rating transition patterns in average stock returns performs better than the three-factor model of Fama and French for a panel of 48 small-cap U.S. entities. Using rolling-average flow rates to derive quarterly cohort transition matrices, we provide evidence to support a statistically significant negative downgrade risk premium in excess returns indicating that stocks at high risk of failure tend to deliver lower returns. The model's performance remains robust across several estimation methods. Whilst panel Granger non-causality tests provide no evidence to support the causal relationship in either direction between excess returns and rating transition probabilities, the basis for further empirical validation and development of the FF-type models under distress intensity are provided.

C1870: **The network effect in credit concentration risk**
Presenter: **Davide Cellai**, Central Bank of Ireland, Ireland
Co-authors: Trevor Fitzpatrick

Measurement and management of credit concentration risk is critical for banks and relevant for micro-prudential requirements. While several methods exist for measuring credit concentration risk within institutions, the systemic effect of different institutions' exposures to the same counterparties has been less explored so far. We propose a measure of the systemic credit concentration risk that arises because of common exposures between different institutions within a financial system. This approach is based on a network model that describes the effect of overlapping portfolios. This metric is applied to synthetic and real world data to illustrate that the effect of common exposures is not fully reflected in single portfolio concentration measures. It also allows us to quantify several aspects of the interplay between interconnectedness and credit risk. Using this network measure, we formulate an analytical approximation for the additional capital requirement corresponding to the systemic risk arising from credit concentration interconnectedness. This methodology also avoids double counting between the granularity adjustment and the common exposure

adjustment. Although approximated, this common exposure adjustment is able to capture, with only two parameters, an aspect of systemic risk that can extend a single portfolios view to a system-wide one.

C1949: **Default rates spillovers: An analysis based on Italian regional data**

Presenter: **Andrea Cipollini**, University of Palermo, Italy

The spatial spillovers mechanism across Italian regions is estimated by using the default rates on loans facilities as a proxy of the loans probability of default, over the period 1996-2015. First, we investigate the presence of spatial dependence across the regional loan default rates. Second, we evaluate whether the Mezzogiorno regions are more affected by spillover effects arising from the Northern regions. For this purpose, we use connectedness measures which are based on the generalized forecast error variance decomposition obtained from the estimation of a Vector Autoregression model. Given the relatively large number of variables, we use the Adaptive elastic net to estimate the VAR model. The empirical findings reveal an increase in default rates spatial dependence overthe 2011Q4 - 2015Q4 (crisis) period, especially for producer households. Moreover, we find evidence of a strong dependence of the Islands from the North of Italy, while the other Southern regions are found to be the most contributor, together with the Northwest of Italy, of financial distress to the remaining gmacro-regions.

Sunday 15.12.2019 14:25 - 16:05 Parallel Session I – CFE-CMStatistics

EI014 Room Beveridge Hall DEPTH, ENSEMBLES AND INFERENCE Chair: Peter Rousseeuw

E0155: **On a generalization and computation of Tukey's depth**
Presenter: **Yiyuan She**, Florida State University, United States

Data depth provides a useful tool for nonparametric statistical inference and estimation but also encounters computational difficulties and scope limitation in modern statistical data analysis. The focus is on the generalization and computation of Tukey's depth for supervised learning in multi-dimensions. Our framework of method-driven halfspace depth emphasizes the importance and properties of the underlying residual space and allows for various distance measures. Moreover, our extension can handle restricted parameter spaces and non-smooth objectives in possibly high dimensions by use of generalized gradients and slack variables. The new formulation of Tukey's depth enables us to utilize state-of-the-art optimization techniques to develop accelerated algorithms with implementation ease and guaranteed fast convergence. Simulations and real data examples demonstrate the efficacy of the proposed methodology in statistical inference and estimation.

E0156: **Accurate parametric inference in high dimensional settings: A step beyond the bootstrap**
Presenter: **Maria-Pia Victoria-Feser**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Mucyo Karemera, Samuel Orso

Accurate estimation and inference in finite sample is important especially when the available data are complex, like when they include mixed types of measurements, they are dependent in several ways, there are missing data, outliers, etc. Indeed, the more complex the data (hence the models), the less accurate are asymptotic theory results in finite samples. This is in particular the case, for example, with logistic regression, with possibly also random effects to account for the dependence structure between the outcomes, or more generally, when the likelihood function or the estimating equations have non closed-form expression. Moreover, resampling techniques such as the Bootstrap can also be quite inaccurate in these settings, unless (complex) corrections are provided. We propose instead a simulation based method, the Iterative Bootstrap (IB), that can be used, very generally, to obtain a) unbiased estimators in high dimensional settings, b) finite sample distributions for inference, with, under suitable conditions, the exact probability coverage property. The method is based on an initial estimator, that does not need to be consistent and can hence be chosen for numerical convenience, and/or can have some desirable properties such as robustness. We present the main theoretical results and the relationships with well-established methods, as well as simulation studies involving complex models and different estimators.

E0157: **Ensemble of regularized linear models**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Ruben Zamar

A new approach is presented for building ensembles of regularized linear models. The approach consists in optimizing an objective function that encourages sparsity within each model and diversity among the models. The procedure works on top of a given penalized linear regression estimator (e.g., Lasso, Elastic Net, SCAD) by fitting the given estimator to possibly overlapping subsets of features, while at the same time encouraging diversity among the subsets, to reduce the correlation between the predictions from each fitted model. The predictions from the models are then aggregated. For the case of an Elastic Net penalty and orthogonal predictors, we give a closed form solution for the regression coefficients in each of the ensembled models. We prove the consistency of our method in possibly high-dimensional linear models, where the number of predictors can increase with the sample size. An extensive simulation study and real-data applications show that the proposed method systematically improves the prediction accuracy of the base linear estimators being ensembled. Possible extensions to GLMs and other models are discussed.

EO176 Room CLO B01 ADVANCED TOOLS FOR FUNCTIONAL AND OBJECT DATA Chair: Hans-Georg Mueller

E1101: **Functional snippets**
Presenter: **Jane-Ling Wang**, University of California Davis, United States

The focus is on the estimation of the mean and the covariance functions of functional snippets, which are short segments of functions possibly observed irregularly on an individual specific subinterval that is much shorter than the entire study interval. Estimation of the covariance function for functional snippets is challenging, since information for the far off-diagonal regions of the covariance structure is completely missing. We address this difficulty by decomposing the covariance function into a variance function component and a correlation function component. The variance function can be effectively estimated nonparametrically, while the correlation part is modeled parametrically, possibly with increasing number of parameters, to handle the missing information in the far off-diagonal regions. Both theoretical analysis and numerical simulations suggest that this hybrid strategy is effective and efficient. In addition, we propose a new estimator for the variance of measurement errors and analyze its asymptotic properties. This estimator is required for the estimation of the variance function from noisy measurements.

E1520: **Additive regression for predictors of various natures and Hilbertian responses with application to censored/missing data**
Presenter: **Byeong Park**, Seoul National University, Korea, South
Co-authors: Ingrid Van Keilegom, Jeong Min Jeon

A fully nonparametric additive regression model for responses and predictors of various natures is considered. This includes the case of Hilbertian and incomplete responses (like censored or missing responses), and continuous, discrete or even nominal predictors. We propose a backfitting technique that estimates this additive model, and establish the existence of the estimator and the convergence of the associated backfitting algorithm under minimal conditions. We also develop a general asymptotic theory for the estimator, which includes even the case where there is no continuous predictor in the model. We verify the practical performance of the proposed estimator in an extensive simulation study, and apply the method to four data sets, containing respectively a missing scalar response, a censored scalar response, a compositional response and a functional response.

E1261: **Partial separability and graphical models for multivariate functional data**
Presenter: **Alexander Petersen**, University of California Santa Barbara, United States
Co-authors: Sang-Yun Oh, Javier Zapata

Graphical models are a ubiquitous tool for identifying dependencies among components of high-dimensional multivariate data. Recently, these tools have been extended to estimate dependencies between components of multivariate functional data by applying multivariate methods to the coefficients of truncated basis expansions. A key difficulty compared to multivariate data is that the covariance operator is compact, and thus not invertible. We will discuss a property called partial separability that circumvents the invertibility issue and identifies the functional graphical model with a countable collection of finite-dimensional graphical models. This representation allows for the development of simple and intuitive estimators. Finally, we will demonstrate the empirical findings of our method through simulation and analysis of functional brain connectivity during a motor task.

E0196: **Bootstrapping max statistics in high dimensions: Near parametric rates under weak variance decay and applications**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Zhenhua Lin, Hans-Georg Mueller

In recent years, bootstrap methods have drawn attention for their ability to approximate the laws of "max statistics" in high-dimensional problems. A

leading example of such a statistic is the coordinate-wise maximum of a sample average of n random vectors in R^p. Existing results for this statistic show that the bootstrap can work when $n << p$, and rates of approximation (in Kolmogorov distance) have been obtained with only logarithmic dependence in p. Nevertheless, one of the challenging aspects of this setting is that established rates tend to scale like $n^{-1/6}$ as a function of n. The main purpose is to demonstrate that improvement in rate is possible when extra model structure is available. Specifically, we show that if the coordinate-wise variances of the observations exhibit decay, then a nearly $n^{-1/2}$ rate can be achieved, independent of p. Furthermore, a surprising aspect of this dimension-free rate is that it holds even when the decay is very weak. Lastly, we provide examples showing how these ideas can be applied to inference problems dealing with functional and multinomial data.

EO584 Room Bloomsbury INSTRUMENTAL VARIABLES METHODS Chair: Babette Brumback

E1234: **The confidence interval method for selecting valid instrumental variables**
Presenter: **Frank Windmeijer**, University of Bristol, United Kingdom
Co-authors: Xiaoran Liang, Fernando Hartwig, Jack Bowden

A new method, the confidence interval (CI) method, is proposed to select valid instruments from a set of potential instruments that may contain invalid ones, for instrumental variables estimation of the causal effect of an exposure on an outcome. Invalid instruments are such that they fail the exclusion restriction and enter the model as explanatory variables. The CI method is based on the confidence intervals of the per instrument causal effects estimates. Each instrument specific causal effect estimate is obtained whilst treating all other instruments as invalid. The CI method selects the largest group with all confidence intervals overlapping with each other as the set of valid instruments. Under a plurality rule, we show that the resulting IV, or two-stage least squares (2SLS) estimator has oracle properties, meaning that it has the same limiting distribution as the oracle 2SLS estimator with the set of invalid instruments known. This result is the same as for the hard thresholding with voting (HT) method. Unlike the HT method, the number of instruments selected as valid by the CI method is guaranteed to be monotonically decreasing for decreasing values of the tuning parameter, which determines the width of the confidence intervals. For the CI method, we can therefore use a downward testing procedure based on the Sargan test for overidentifying restrictions.

E1295: **Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of pleiotropy**
Presenter: **Jack Bowden**, University of Exeter, United Kingdom

Mendelian randomization (MR) uses genetic variants as instrumental variables to estimate the causal effect of a modifiable health exposure on a downstream health outcome. The technique can be implemented using only summary data estimates of genetic association from genome wide association studies, which is referred to as 'two sample summary data MR. Typically many hundreds of genetic are used in such an analysis, but it is highly likely that a sizable number of them are, in fact, invalid instruments. A main cause of instrument invalidity is when a genetic variant exerts a direct effect on an outcome not through the exposure of interest. This is a violation of the exclusion restriction, and is referred to in the MR field as 'pleiotropy'. We develop a Bayesian Model Averaging approach that intelligently searches the space of all 2^L models (L being the number of genetic variants) to obtain a more robust and reliable causal estimate. Our algorithm favours models with large numbers of variants, but down-weights sets of variants that lead to heterogeneous causal effect estimates. It also naturally accounts for weak instrument bias via the use of a a posterior profile likelihood function. We illustrate our approach first for a basic one parameter causal model, and then show how it can be extended to more complex modelling frameworks.

E1365: **Three causal lessons from a simulation learner: On SUTVA, instrumental variables and causal estimands**
Presenter: **Els Goetghebeur**, Ghent University, Belgium
Co-authors: Saskia le Cessie

To support Causal reasoning from DAGS and a choice among available estimators, the STRATOS causal inference group developed a 'simulation learner'. This engine generates per subject alongside observed exposure(s) and outcome a range of alternative exposures with their potential outcome. As in the Promotion of Breastfeeding Intervention Trial, we 'randomize mother-infant pairs to standard of care or a breastfeeding encouragement (BFE) intervention. Main outcome is weight at 3 months. The path from randomization to outcome meets the intervention uptake (education program), followed by the start and a specific duration of breastfeeding. Simulated parallel worlds then enable visualization of various potential outcomes and causal estimands in specific populations. The necessary intermediate steps highlight that SUTVA must be context specific. We see randomisation act as an instrument for one exposure (e.g receiving an offer for the BFE programme or actually following the BFE programme), but not others (e.g. actually starting breastfeeding). We recognize that averaging causal effects over an observed (experimental) instrument may be irrelevant unless one conditions on the instrumental variable. We thus explore distinct estimation methods and compare results with the simulated population parameters. R code is available on www.ofcaus.org, where SAS and Stata code for analysis is also provided This is work on behalf of STRATOS TG 7 Causal Inference

E1415: **Structural nested models for cluster-randomized trials with cluster-level non-adherence**
Presenter: **Babette Brumback**, University of Florida, United States

Much attention has been paid to estimating the causal effect of adherence to a randomized protocol using instrumental variables to adjust for unmeasured confounding. The interest stems from a wish to estimate the effect of cluster-level adherence on individual-level binary outcomes with a three-armed cluster-randomized trial and polytomous adherence. We developed two structural-nested modeling approaches for estimation; the approaches differ in the handling of measured individual-level confounders of the effect of randomization on the outcome. The first approach uses a weighted generalized structural nested mean model, which adjusts for the confounders using weights, and the second approach uses an ordinary generalized structural nested mean model, which stratifies on the confounders. The two approaches target different estimands. The methodology accommodates cluster-randomized trials with unequal probability of selecting individuals. Furthermore, we developed a method to implement the approaches with relatively simple programming. The approaches work reasonably well, but when the structural-nested model does not fit the data, there may be no solution to the estimating equation. We investigate the performance of the approaches using simulated data, and we also use them to estimate the effect on pupil absence of school-level adherence to a randomized water, sanitation, and hygiene intervention in western Kenya.

EO192 Room G11 ADVANCES IN STATISTICAL NETWORK ANALYSIS Chair: Jonathan Stewart

E0615: **Spectral inference for large stochastic blockmodels with nodal covariates**
Presenter: **Angelo Mele**, Johns Hopkins University, United States
Co-authors: Joshua Cape, Carey Priebe, Lingxin Hao

Spectral methods are studied for inference in large stochastic blockmodels with observed nodal covariates. We formulate the estimation problem as recovery of latent positions in the Generalized Random Dot Product Graph (GRDPG) model, thereby extending recent advances in spectral methods to provide an algorithm that simultaneously estimates the block assignments and parameters for observed covariates. The spectral estimator is asymptotically normal and computationally fast, when compared to a standard variational EM algorithm. The results provide a foundation to estimate the effect of observed covariates as well as unobserved latent community structure on the probability of link formation in massive networks.

E1180: **Modeling the dynamics of social network perceptions**
Presenter: **Nynke Niezink**, Carnegie Mellon University, United States

To understand and predict the behavioral consequences of social networks, it is important to understand how social networks form. Studies of

network dynamics usually rely on data of the network ties (e.g., friendship or collaboration) among a group of social actors, such as people or organizations, collected at multiple measurement moments. Many studies have shown that individuals differ in how they perceive and cognitively represent the networks they are embedded in. However, in the analysis of network dynamics, this is not taken into account. Instead, current models assume individuals to make network decisions, creating and dissolving ties, based on a shared network representation. We propose a model for the dynamics of social networks taking individuals network perceptions into account. This model generalizes the stochastic actor-oriented model, a continuous-time Markov chain model on the state space of all possible networks among a group of actors, to simultaneously model the network as perceived by all actors.

E0922: **Generalized beta-models with dependent edges and parameter vectors of increasing dimension**
Presenter: **Jonathan Stewart**, Rice University, United States
Co-authors: Michael Schweinberger

An important question in statistical network analysis is how to construct random graph models with dependent edges without sacrificing computational scalability and statistical guarantees. We advance models, methods, and theory by introducing a probabilistic framework for populations consisting of overlapping subpopulations of similar or dissimilar sizes, which allows dependence to propagate throughout the population graph. As examples, we introduce generalizations of beta-models with dependent edges capturing brokerage in social networks. On the statistical side, we derive consistency results in settings where dependence propagates throughout the population graph, and the number of parameters increases with the number of subpopulations. We show how the rate of convergence depends on the amount of overlap and the sizes of subpopulations, how different the sizes are, and how sparse the population graph is. On the computational side, we demonstrate how the conditional independence structure of models can be exploited for local computing.

E1838: **Nonparametric estimation for multiple heterogeneous networks**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom
Co-authors: Pierre-Andre Maugis

Nonparametric estimation is studied for the setting where multiple networks are observed on the same set of entities (nodes), with or without covariate information. Such samples may arise in the form of replicated networks assumed to be drawn from a common distribution, or in the form of longitudinal networks observed over time or space with the network generating process varying from one network to another. For example, social interaction networks between subjects over time or on different social media platforms; in connectomics where a brain network is observed for each subject along with age, gender etc. Drawing on concepts and techniques from graph theory and embedding approaches, we show how standard nonparametric methods can be employed to lead to a simple kernel estimator. Unlike existing histogram or blockmodel approximations to graphon function, our method allows estimation of node-specific as well as network-specific heterogeneity and hence offers an easy to interpret and flexible approach.

EO266 Room G3 STATISTICAL METHODS APPLIED TO INSURANCE AND ACTUARIAL SCIENCES Chair: Olivier Lopez

E0183: **A micro-level study of IBNR claims reporting delays using extreme value theory**
Presenter: **Maud Thomas**, Sorbonne University, France
Co-authors: Jonathan El Methni

The evaluation of the volume of IBNR claims (Incurred But Not Reported) is a challenging task in claim reserving. A standard way to proceed is to rely on chain-ladder type techniques. These techniques are based on an aggregate vision of the risk, and on a stability of the payment process. The prediction obtained via chain-ladder usually does not distinguish between IBNR and RBNS claims (Reported But Not Settled). Recently, interest in looking more precisely in the reporting dynamic has increased. We propose a method to perform a closer look of IBNR by studying the distribution of a large delay before reporting a claim at a micro-level using Extreme Value Theory. The distribution of largest IBNR claims reporting delays belongs to the family of Weibull-tail distributions. Such distributions have already been used to model large claims in non-life insurance. The behaviour of these distributions is characterised by a shape parameter, called the Weibull-tail coefficient. We derive a data-driven procedure to estimate this coefficient using techniques inspired by Lepski's method.

E0252: **Micro forecasting**
Presenter: **Michal Pesta**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Matus Maciak, Ostap Okhrin

Forecasting costs is now a front burner in empirical economics. We propose an unconventional tool for stochastic prediction of future expenses based on the individual (micro) developments of recorded events. Let us think of a firm, enterprise, institution, or state, which possesses knowledge about particular historical events. For each event, there are several related payments or losses spread over time. Nevertheless, the issue is that some already occurred events do not have to be necessarily reported. The aim lies in forecasting future cash flows coming from already reported, occurred but not reported, and yet not occurred events. Our methodology is illustrated on quantitative risk assessment, however, it can be applied to other areas such as startups, epidemics, war damages, advertising and commercials, digital payments, or drug prescription as discussed in the paper. As a theoretical contribution, stochastic inference for marked non-homogeneous Poisson process with non-homogeneous Poisson processes as marks is developed.

E0576: **Nonparametric copula estimation for mixed insurance claim data**
Presenter: **Lu Yang**, University of Amsterdam, Netherlands

Multivariate claim data are common in insurance applications, e.g. claims of each policyholder from different insurance coverages. Understanding the dependencies of such multivariate risks is essential for the solvency and profitability of insurers. However, at the policyholder level, claim outcomes usually follow a hybrid distribution with a large point mass at zero corresponding to the case of no claims, while some customers report positive claims. In order to accommodate complex features of the marginal distributions while flexibly quantifying the dependencies among multivariate claims, we employ copulas. Although a substantial literature focusing on copula models with continuous outcomes has emerged, some key steps do not carry over to mixed data. In particular, existing nonparametric copula estimators are not consistent for mixed data. Thus, copula specification and diagnostics with mixed outcomes has remained a problem. We fill in this gap by developing a nonparametric copula estimator for mixed data. We show the uniform convergence of the proposed nonparametric copula estimator, and through simulation studies, we demonstrate that the probabilities of zero play a crucial role for the finite sample performance of the proposed estimator. Using the claim data from the Wisconsin Local Government Property Insurance Fund, we illustrate that our nonparametric copula estimator can assist analysts in identifying important features of the underlying dependence structure.

E0680: **Cyber claim analysis through generalized Pareto regression trees with applications to insurance pricing and reserving**
Presenter: **Sebastien Farkas**, Sorbonne Universite, CNRS, Laboratoire de Probabilites, Statistique et Modelisation, LPSM, France
Co-authors: Olivier Lopez, Maud Thomas

Cyber claim databases are heterogeneous and contain extreme values. This heterogeneity is caused by the evolution of the risk but also by the evolution in the quality of data and of sources of information through time. We propose a methodology to analyze the heterogeneity of cyber claim databases using regression trees. We consider a public database considered as a benchmark for cyber event and more specifically for data breaches. Particular attention is paid to the tail of the distribution, using a generalized Pareto likelihood as splitting criterion for growing the regression tree. Combining this analysis with, on the one hand, a model for the frequency of the claims, and on the other hand, a model for loss quantification of

data breaches, we develop a simple model for pricing and reserving in cyber insurance.

EO164 Room G5 RECENT ADVANCES IN FUNCTIONAL AND MULTIVARIATE DATA ANALYSIS Chair: Yuko Araki

E1025: **Time course modeling for brain imaging data**
Presenter: **Atsushi Kawaguchi**, Saga University, Japan

Brain time varying information is useful for identifying biomarkers that can be used for diagnosis of brain disorders. This can be measured as longitudinal or time series Magnetic Resonance Imaging (MRI) data. We propose a dimension-reduction method using supervised (multi-block) sparse component analysis. The method is first implemented through basis expansion of spatial brain images, and the scores are then reduced through regularized matrix decomposition to produce simultaneous data-driven selections of related brain regions, supervised by univariate composite scores representing linear combinations of covariates. Two advantages of the proposed method are that it identifies the associations between brain regions at the voxel level, and that supervision is helpful for interpretation. This also regards the functional data analysis approach, which can be applied to the time course modeling. The proposed method was applied to the real data and was compared with the existing methods.

E1074: **Joint modeling and estimation for multivariate longitudinal data and binary outcome**
Presenter: **Toshihiro Misumi**, Yokohama City University, Japan
Co-authors: Hidetoshi Matsui

In medical research areas, a joint modeling that simultaneously incorporates multivariate longitudinal biomarker processes and a binary outcome process has attracted considerable attention. The joint model consists of a multivariate linear mixed effects model and a logistic regression model with shared random effects. Numerous unknown parameters included in the two submodels are simultaneously estimated by joint maximum likelihood method. We discuss the effective estimation procedure based on a pseudo-likelihood and a h-likelihood. The estimated joint model provides a powerful tool to know how closely the multivariate longitudinal trajectories of biomarkers are associated with a clinical outcome. We also discuss the relationship between the joint modeling and functional data modeling. Some numerical studies are conducted to investigate the effectiveness of our proposed modeling strategy.

E1107: **Causal inference in imaging genetics**
Presenter: **Dehan Kong**, University of Toronto, Canada

Understanding the workings of human brains and their connections with dementia behaviour is a central goal in medical studies. We will introduce a new method to identify the causal relationship between hippocampal atrophy and dementia behaviour in Alzheimer disease. We consider a 2D hippocampal surface exposure and develop a causal inference procedure which can account for high dimensional potential genetic confounders. We examine the performance of our method using a large-scale imaging genetic dataset from the Alzheimer Disease Neuroimaging Initiative study.

E1270: **Statistical inference on M-estimators by high-dimensional Gaussian approximation**
Presenter: **Masaaki Imaizumi**, The Institute of Statistical Mathematics, Japan

A statistical inference method is developed for a general class of estimators with fewer restrictions. Measuring the uncertainty of estimators, such as asymptotic normality, is a fundamental and standard tool for statistical inference such as a statistical test and a confidence analysis. However, there are several situations that we cannot evaluate its uncertainty, for example, non-differentiable loss functions and parameter spaces as the non-Donsker class. We consider an M-estimator which is defined as an argmax of an empirical mean of criteria functions. Then, we approximate a distribution of the M-estimator by a supremum of a known Gaussian process. For the method, we employ a notion of the high-dimensional Gaussian approximation and apply it to the approximation. We provide a theoretical bound for an error of the approximation. Moreover, we propose a multiplier bootstrap method for statistical inference.

EO310 Room MAL G13 BAYESIAN APPLICATIONS AND METHODS Chair: Christopher Hans

E0407: **Modeling crime dynamics and associations with the built environment in Philadelphia**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States

Urban data analysis has been recently improved through publicly available high resolution data, allowing us to empirically investigate urban design principles of the past half century. We will focus on the spatial-temporal modeling of the change in crime over the past decade in the city of Philadelphia. We will explore different parametric and non-parametric Bayesian approaches for finding regions of the city that share similar crime dynamics. Within this context, we develop methodology for non-parametric clustering of regions simultaneously across multiple levels of spatial resolution. We will also investigate whether crime in particular locations is associated with aspects of the built environment, such as different types of land use or presence of different types of businesses.

E0450: **Evaluating change in learning from different forms of interactive visualizations with a large case study**
Presenter: **Leanna House**, Virginia Tech, United States
Co-authors: Lata Kodali

Cutting edge software has been preciously developed that allows novice analysts to explore high-dimensional data interactively. The software, Andromeda, effectively responds to user inputs in the forms of interaction with data at the observation-level and/or parametric levels to create multiple Weighted Multidimensional Scaling (WMDS) projections. We evaluate the impact Andromeda has on student learning via a large-scale user study implemented in an introductory statistics course at Virginia Tech. This study includes approximately 150 students and was conducted in two different semesters. Using a Bayesian approach, we share our findings from this user study, including significant differences in mastery of WMDS, complexity of insights, and change in attitude toward engaging in data analyses.

E1306: **Fast computing for latent Gaussian random field models**
Presenter: **Murali Haran**, The Pennsylvania State University, United States

Latent Gaussian random field models are extremely popular in a wide variety of areas, including environmental science and infectious disease modeling. We will describe some computational efficient strategies for fitting such models within a Bayesian paradigm. Our approach applies to a wide array of latent Gaussian random field models, and permits the analysis of large data sets.

E1313: **Community detection in co-location networks**
Presenter: **Catherine Calder**, University of Texas, Austin, United States

The extent to which activity spaces - the collection of an individuals routine activity locations - overlap provides important information about the function of a city and its neighborhoods. To study the patterns of overlapping activity spaces and to detect communities of individuals based on their shared locations, we introduce the notion of an ecological network, a type of two-mode network with the two modes being individuals and routine activity locations. Specifically, we identify latent activity pattern profiles, which, for each community, summarize its members probability distribution of going to each location, and community assignment vectors, which, for each individual, summarize his/her probability distribution of belonging to each community. Using data from the Adolescent Health and Development in Context (AHDC) Study, we employ Bayesian probabilistic topic models to identify activity pattern profiles and community assignment vectors. We then explore differences across neighborhoods of Columbus, OH in the strength, and within-neighborhood consistency of community assignment, paying particular attention to the association

between race and these measures.

EO128 Room MAL G14 ADVANCES IN BAYESIAN COMPUTATION Chair: Radu Craiu

E0448: **Efficient Bernoulli factory MCMC for intractable likelihoods**
Presenter: **Dootika Vats**, Indian Institute of Technology, Kanpur, India

Accept-reject based Markov chain Monte Carlo (MCMC) algorithms have traditionally been a function of the ratio of the target density at two contested points. We note that this feature is rendered almost useless in Bayesian MCMC problems within tractable likelihoods. We introduce a new acceptance probability that has the distinguishing feature of not being as a function of the ratio of the target density at two points. We show that such a structure allows for the construction of an efficient and stable Bernoulli factory. The resulting "Portkey Barker's" algorithm is exact and is computationally more efficient that the current state-of-the-art.

E0544: **On the scalability of conditional particle filters**
Presenter: **Matti Vihola**, University of Jyvaskyla, Finland
Co-authors: Anthony Lee, Sumeetpal Singh

Hidden Markov models (HMMs) are a flexible framework for time-series modelling. Full Bayesian inference of non-linear and/or non-Gaussian HMMs has remained a challenge until the recently introduced particle Markov chain Monte Carlo methods. In particular, the conditional particle filter (CPF), and its backward sampling variant (CBPF), have been found efficient in many challenging settings. We discuss the scalability properties of the CPF and the CBPF with respect to the time horizon (length of the time series). Our theoretical results align well with the empirical observations about the efficiency. In particular, our findings about the CBPF confirm the long held view that the CBPF remains an effective sampler with a fixed number of samples even as the time horizon increases. Our analysis of the CBPF relies on analysis of a so-called coupled conditional backward sampling particle filter (CCBPF) algorithm, which is interesting on its own right. Indeed, CCBPF is a simple algorithmic variant of PREVIOUS methods suggested for unbiased estimation with respect to the smoothing distribution of a HMM.

E1003: **A long short-term memory stochastic volatility model**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: Minh-Ngoc Tran, David Gunawan, Nghia Nguyen

Stochastic Volatility (SV) models are widely used in the financial sector while Long Short-Term Memory (LSTM) models have been successfully used in many large-scale industrial applications of Deep Learning. Our article combines these two methods non trivially and proposes a model for capturing the dynamics of financial volatility process, which we call the LSTM-SV model. The proposed model overcomes the short-term memory problem in conventional SV models, is able to capture non-linear dependence in the latent volatility process, and often has a better out-of-sample forecast performance than SV models. The conclusions are illustrated through simulation studies and applications to three financial time series datasets: US stock market weekly index SP500, Australian stock weekly index ASX200 and Australian-US dollar daily exchange rates. We argue that there are significant differences in the underlying dynamics between the volatility process of SP500 and ASX200 datasets and that of the exchange rate dataset. For the stock index data, there is strong evidence of long-term memory and non-linear dependence in the volatility process, while this is not the case for the exchange rates.

E1638: **Convergence time of some non-reversible MCMC methods**
Presenter: **Florian Maire**, universite de montreal, Canada

It is commonly admitted that non-reversible Markov chain Monte Carlo (MCMC) algorithms usually yield more accurate MCMC estimators than their reversible counterparts. Some recent results have established an ordering showing that a large class of non-reversible algorithms tend to have a smaller asymptotic variance than their reversible equivalent. In particular, the Guided Walk or some lifting techniques are shown to have a smaller asymptotic variance than the Metropolis algorithm. We show that in addition to their variance reduction effect, some non-reversible MCMC algorithms have also the undesirable property to slow down the convergence of the Markov chain towards its stationary distribution. This point, which has been overlooked by the literature, has obvious practical implications. We illustrate this small asymptotic variance/slow convergence scenario phenomenon for different non-reversible versions of the Metropolis algorithm on several discrete state space examples. Our findings echo an important discussion related to the design of the refreshment rate in several non-reversible algorithms including the bouncy particle sampler: Markov chains that are too irreversible see their rate of convergence slowing down and a trade-off is required. We present simple adjustments of some non-reversible MCMC algorithms to mitigate this risk.

EO338 Room MAL G15 TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Bernardo Nipoti

E0625: **A simultaneous transformation and rounding approach for modeling integer-valued data**
Presenter: **Daniel Kowal**, Rice University, United States
Co-authors: Antonio Canale

A simple yet powerful framework for modeling integer-valued data is proposed. The integer-valued data are modeled by Simultaneously Transforming And Rounding (STAR) a continuous-valued process, where the transformation may be known or learned from the data. Implicitly, STAR formalizes the commonly-applied yet incoherent procedure of (i) transforming integer-valued data and subsequently (ii) modeling the transformed data using Gaussian models. Importantly, STAR is well-defined for integer-valued data, which is reflected in predictive accuracy, and is designed to account for zero-inflation, bounded or censored data, and over- or underdispersion. Efficient computation is available via an MCMC algorithm, which provides a mechanism for direct adaptation of successful Bayesian methods for continuous data to the integer-valued data setting. Using the STAR framework, we develop new linear regression models, additive models, and Bayesian Additive Regression Trees (BART) for integer-valued data, which demonstrate substantial improvements in performance relative to existing regression models for a variety of simulated and real datasets.

E1047: **Bayesian nonparametric dynamic clustering: An application to gender stereotypes in words**
Presenter: **Alessandra Guglielmi**, Politecnico di Milano, Italy
Co-authors: Maria De Iorio, Stefano Favaro, Lifeng Ye

A probability model is proposed for a collection of random distributions indexed by time. The model is based on the dependent Dirichlet process prior and dependence among the random measures is introduced via latent variables, through a Gaussian copula transformation of a stationary autoregressive process of order one. This allows us to introduce time dependence among the random distributions. We propose a Sequential Monte Carlo algorithm to perform posterior inference in case of random densities given by mixtures of this time-dependent family of a.s. discrete distributions. We apply this model to a very interesting problem, to understand how gender stereotypes in words had changed over time in the 20th century. Other typical applications involve multiple time series in the biomedical context, as well as population density dynamics with areal data. Advantages of the proposed approach include wide applicability, ease of computations, interpretability and time dependent clustering of the observation. K-step nonparametric predictive density functions can be derived. The model retains desirable statistical properties for inference, while achieving substantial flexibility.

E1361: **Weighted dynamic multi-layer networks via latent Gaussian processes**
Presenter: **Christian Carmona**, University of Oxford, United Kingdom
Co-authors: Serafin Martinez-Jaramillo

A general network model suited for longitudinal data of multi-layer networks with directed and weighted edges is proposed. The formulation combines relevant features from existing network models, creating one that is able to capture simultaneously the characteristics of such complex networks. The model is built upon the *latent social space* representation of networks. It consists of a hierarchical formulation: deep levels of the model represent latent coordinates of agents in the social space, evolving in continuous time via Gaussian processes; meanwhile, top levels jointly manage incidence and strength of interactions by considering a mixture between a Gaussian component and a point-mass probability at zero. Learning of the model is performed through Bayesian Inference. We develop an efficient MCMC algorithm targeting the posterior distribution of model parameters and missing data (code available in the supplement). The performance of the model is measured in synthetic data, as well as our main case study: the network of inter-bank transactions in the Mexican financial system. Accurate predictions are obtained in both cases estimating out-of-sample link incidence and link strength.

E0596: **Distributed Bayesian inference for varying coefficient spatiotemporal models**
Presenter: **Cheng Li**, National University of Singapore, Singapore
Co-authors: Rajarshi Guhaniyogi, Terrance Savitsky, Sanvesh Srivastava

Bayesian varying coefficients modeling is popular in many disciplines due to its flexibility and interpretability. Markov chain Monte Carlo methods used to fit these models are inefficient in moderately large data. We address this problem by developing a generalization of this class of models using linear mixed effects modeling, where the random effects are modeled by Gaussian processes. Computationally, we use parameter expansion to develop an efficient and stable data augmentation-type algorithm for fitting these models under the Bayesian framework, which can be scaled to millions of observations using the divide-and-conquer technique. Theoretically, we derive the convergence rates of Bayes risks for the divide-and-conquer posterior distributions of parameters, and show that the rates can be tuned to nearly optimal when the true underlying function is assumed to lie in some general functional classes. We demonstrate that our method yields smaller mean square errors, shorter credible intervals, and better frequentist coverage for the model parameters than its competitors, using several numerical experiments and real data applications.

EO392 Room MAL G16 RECENT DEVELOPMENTS IN VINE COPULAS Chair: Thomas Nagler

E0742: **Vine copula autoencoders**
Presenter: **Thibault Vatter**, Columbia University, United States
Co-authors: Natasha Tagasovska, Damien Ackerer

A vine copula autoencoder is proposed to construct flexible generative models for high-dimensional distributions in a straightforward three-step procedure. First, an autoencoder compresses the data using a lower dimensional representation. Second, the multivariate distribution of the encoded data is estimated with vine copulas. Third, a generative model is obtained by combining the estimated distribution with the decoder part of the autoencoder. This approach can transform any already trained autoencoder into a flexible generative model at a low computational cost. This is an advantage over existing generative models such as adversarial networks and variational autoencoders which can be difficult to train or impose strong assumptions on the latent space. Experiments on MNIST, Street View House Numbers and Large-Scale CelebFaces Attributes datasets show that vine copulas autoencoders achieve competitive results.

E0744: **Vine models for stationary time series**
Presenter: **Thomas Nagler**, Leiden University, Germany
Co-authors: Aleksey Min

Vine copulas can be used to simultaneously model inter-serial and cross-sectional dependence in stationary time series. We introduce a class of vine structures that is particularly suited for such applications. The class is characterized in a way that stationarity is guaranteed if pair-copulas are invariant to shifts in the time index (which is a necessary condition in any case). The class includes several structures that were already proposed in the literature.

E0755: **Dynamic regular vine copulas with an application to exchange rates dependence**
Presenter: **Alexander Kreuzer**, Technische Universitat Munchen, Germany

Modeling dependence among financial assets is an important research topic, since the dependence structure has high influence on the risk associated with a corresponding portfolio. Regular vine copulas have proven as a useful tool in this context. They allow for characteristics like asymmetric tail dependence, which cannot be modeled with a multivariate Gaussian or Student-t copula. Usually it is assumed that the dependence parameters of the regular vine copula remain constant as time evolves. We get rid of this assumption and propose dynamic regular vine copulas. In this dynamic model dependence parameters are described through latent AR(1) processes. Since maximum likelihood estimation is infeasible for these latent AR(1) processes, we employ Markov Chain Monte Carlo within a sequential estimation procedure. The approach is illustrated with 25-dimensional exchange rates data, where we find clear evidence for dynamic dependence.

E1714: **A new algorithm for finding a good fitting truncated R-vine copula in high dimensions**
Presenter: **Edith Alice Kovacs**, University of Debrecen, Hungary
Co-authors: Tamas Szantai

Modelling multivariate probability distributions by using R-vine copulas gained popularity due to their flexibility in modelling many types of dependences in the same time. However, in high dimensions, this is also their main drawback, because they involve a large number of parameters. To tackle this problem, two main approaches were proposed in the literature, namely, the truncation and the simplification of the vine copulas. We present a new algorithm for fitting truncated vines. The basic idea is the exploration of the conditional independences defining the truncated vine copula structure. We will show some advantages of the proposed method.

EO062 Room CLO 101 SUBSAMPLING METHODS FOR MASSIVE DATA Chair: HaiYing Wang

E0454: **Statistical inference in big data high dimensional generalized estimating equations via optimal subsampling**
Presenter: **Hanxiang Peng**, IUPUI, United States

Statistical inference is considered for massive data sets in the framework of general estimating equations when the sample size is massive. Our approach is the A-optimal subsampling. We derive the A-optimal sampling distributions, discuss the algorithmic and statistical properties, construct the A-optimal Scoring Algorithm, and provide the asymptotic behaviors of the subsampling estimates for both fixed and growing dimension. In the end, we report some simulations and real data results.

E0823: **Optimal subsampling: Sampling with replacement vs Poisson sampling**
Presenter: **HaiYing Wang**, University of Connecticut, United States

Faced with massive data, subsampling is a commonly used technique to improve computational efficiency, and using nonuniform subsampling probabilities is an effective approach to improve estimation efficiency. In the context of maximizing a general target function, optimal subsampling probabilities are derived for both subsampling with replacement and Poisson subsampling. The optimal subsampling probabilities minimize functions of the subsampling approximation variances in order to improve the estimation efficiency. Furthermore, they provide deep insights on the theoretical similarities and differences between subsampling with replacement and Poisson subsampling. Practically implementable algorithms are proposed based on the optimal structural results, which are evaluated by both theoretical and empirical analysis.

E0991: **Design based incomplete U-statistics**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: Xiangshun Kong

U-statistics are widely used in the fields of economy, machine learning and statistics. While it enjoys very desirable statistical properties, it also admits an obvious drawback: the computation easily becomes impractical as the data size n increases. Particularly, the number of combinations, say m, that a U-statistic of order d has to evaluate is in the order of $O(n^d)$. Many efforts have been made to approximate the original U-statistic by a small subset of the combinations in history. Such an approximation was coined as an incomplete U-statistic. To the best of our knowledge, all existing methods require m to grow at least faster than n, albeit much slower than n^d, in order for the corresponding incomplete U-statistic to be asymptotically efficient in the sense of mean squared error. We introduce a new type of incomplete U-statistics, which can be asymptotically efficient even when m grows slower than n. In some cases, m is only required to grow faster than $\sqrt{n}$. Both theoretical and empirical results show significant improvements on the statistical efficiency by the new incomplete U-statistic.

E2000: **Sequential online subsampling for thinning experimental designs**
Presenter: **Luc Pronzato**, CNRS - Universite Cote d'Azur, France
Co-authors: HaiYing Wang

In the considered design problem, experimental conditions (design points X_i) are presented in the form of a sequence of i.i.d. random variables, generated with an unknown probability measure μ, and only a given proportion $\alpha \in (0,1)$ can be accepted. The objective is to select good candidates X_i on the fly and maximise a concave function Φ of the information matrix. The optimal solution corresponds to the construction of an optimal bounded design measure $\xi^*_\alpha \leq \mu/\alpha$. The difficulty is that μ is unknown and ξ^*_α must be constructed online. The construction proposed relies on the definition of a threshold τ on the directional derivative of Φ at the current information matrix, the value of τ being fixed by a certain quantile of the distribution of this directional derivative. Combination with recursive quantile estimation yields a nonlinear two-time-scale stochastic approximation method. It can be applied to very long design sequences, since only the current information matrix and estimated quantile need to be stored. Convergence to an optimum design is proved. Various illustrative examples are presented.

EO096 Room CLO 102 COMPLEX DATA IN THEORY AND PRACTICE Chair: Johannes Lederer

E0455: **Statistical assessment of depth normalization methods for microRNA sequencing**
Presenter: **Li-Xuan Qin**, Memorial Sloan Kettering Cancer Center, United States

Quality data is the foundational cornerstone for reliable scientific findings in evidence-based medical research. It is widely accepted that a crucial step to derive high-quality genomics data is to identify data artifacts caused by systematic differences in the processing of specimens and to remove these artifacts by data normalization. Statistical methods for normalizing sequencing data depth have been recently developed, including re-scaling-based and regression-based methods. Many of these methods rely on the presupposition that variations in the assumed scaling factor or projection of the assumed regression function are solely due to data artifacts and should be removed. MicroRNAs are a unique class of small RNAs closely linked to carcinogenesis. They are low-complexity molecules that tend to be expressed in a tissue-specific manner. As a result, the assumption of depth normalization methods may not hold for microRNA sequencing. We performed a study to assess the performance of existing depth normalization methods on identifying disease relevant microRNAs using both a pair of datasets on the same set of tumor samples and data simulated from this dataset pair under various scenarios of differential expression. We will report our findings from this study.

E0698: **Penalized angular regression for personalized predictions**
Presenter: **Kristoffer Hellton**, Norwegian Computing Center, Norway

A novel penalized regression method is introduced which is specifically constructed to personalize predictions. Personalized angle (PAN) regression estimates a covariate vector-specific regression coefficients, shrinking them in terms of their angles utilizing a hyperspherical parametrization. It is shown that the PAN estimate will be the solution of a low-dimensional eigenvector problem, which for an orthonormal design matrix has an explicit solution. We prove that by combining the PAN and the L2 penalty the resulting prediction will have uniformly smaller asymptotic mean squared error than ridge regression. The resulting estimator is illustrated in a medical application.

E0286: **Backfitting tests in generalised structured models**
Presenter: **Stefan Sperlich**, University of Geneva, Switzerland
Co-authors: Enno Mammen

Bootstrap tests for generalised structured models are introduced. They can be applied for testing interaction terms or the impact of certain covariates in a large family of semiparametric models. This implies, for example, additivity testing, variable selection, linearity, constant returns or checking for endogeneity. The test is based on a comparison of non- and semiparametric alternatives, i.e both the null hypothesis and the alternative are non- or semiparametric. The test statistic is a weighted L2-distance of the two fits, where both estimation procedures make use of the smooth backfitting technique. Asymptotic theory for the test is developed, implementation and computational issues are discussed. Simulations and comparison studies show an excellent performance of the test procedures. We apply the method to additive and varying coefficient models.

E1975: **Domain adaptation from a pre-trained source model**
Presenter: **Luxin Zhang**, Inria/Worldline, France
Co-authors: Pascal Germain, Christophe Biernacki, Yacine Kessaci

Traditional statistical learning paradigm assumes the consistency between train and test data distributions. This rarely holds in many real-life applications. The domain adaptation paradigm proposes a variety of techniques to overcome this issue. Most of the works in this area seek either for a latent space where source and target data share the same distribution, or for a transformation of the source distribution to match the target one. Both strategies require learning a model on the transformed source data. An original scenario is studied where one is given a model that has been constructed using expertise on the source data that is not accessible anymore. To use directly this model on target data, we propose to learn a transformation from the target domain to the source domain. Up to our knowledge, this is a new perspective on domain adaptation. This learning problem is introduced and formalized. We study the assumptions and the sufficient conditions mandatory to guarantee a good accuracy when using the source model directly on transformed target data. By pursuing this idea, a new domain adaptation method based on optimal transport is proposed. We experiment our method on a fraud detection problem.

EO659 Room Court PREPARING FOR THE FUTURE: PHD PROGRAMS IN STATISTICS EDUCATION Chair: Erin Blankenship

E0660: **Ph.D. programs in statistics education: Department of mathematical sciences perspective**
Presenter: **Jennifer Green**, Montana State University, United States

Departments of Mathematical Sciences have traditionally offered Ph.D. programs in Statistics and Mathematics, but what does a Ph.D. program in Statistics with a concentration in Statistics Education encompass? We will discuss the program we recently developed at Montana State University, highlighting the required coursework and how it compares to the traditional Ph.D. program in Statistics. We will also share some of the teaching and research opportunities students have had and describe their dissertation research that encompass a variety of topics, including graduate teaching assistants' development and motivation to teach, pre-service elementary mathematics teachers' statistical knowledge for teaching probability, and graduate students' development and use of statistical computing for research in the environmental sciences.

E0925: **The University of Minnesota doctoral program in statistics education research**
Presenter: **Robert delMas**, University of Minnesota, United States

Statistics education research represents an area of emphasis within the Quantitative Methods in Education (QME) doctoral program of the Department of Educational Psychology at the University of Minnesota. All students who receive a doctoral degree in Educational Psychology complete core coursework in learning theory, cognitive psychology, social psychology, and both quantitative and qualitative educational research methods. In addition, the core requirements of the QME program focus on measurement with additional coursework in applied statistics and evaluation methods. Students who pursue an emphasis in statistics education research gain extensive experience in applying research-based methods for teaching statistics and complete doctoral studies that explore assessment development, curriculum development, problem solving, and conceptual development related to understanding statistics. The coursework, teaching, and research experiences of students in the statistics education research emphasis are detailed. Examples of the research conducted by students are provided, and the types of positions currently held by graduates of the program are indicated.

E1089: **Using funding and collaborations with statistics departments to build interest in statistics education**
Presenter: **Tim Jacobbe**, Southern Methodist University, United States

Statistics Education programs at the University of Florida and now Southern Methodist University have been built through collaborations with statistics departments. Masters students in statistics who are interested in and passionate about the teaching and learning of statistics at the secondary or undergraduate levels need opportunities other than pursuing careers as research statisticians. Students who are interested in statistics education are able to contribute to the growing need for faculty and researchers that are interested in the teaching and learning of statistics. The doctoral student experience as well as discuss positions graduates of the program have filled over the years will be detailed. There is a clear need and an exciting opportunity to bridge programs across department of statistics and schools of education.

E0910: **A graduate student's perspective of a PhD program in statistics education**
Presenter: **Ella Burnham**, University of Nebraska- Lincoln, United States

I will share my experience as a Statistics PhD student with an emphasis in statistics education at the University of Nebraska-Lincoln, including the coursework, teaching experiences, and research experiences I have had compared to other graduate students in the program. More specifically, I will discuss the opportunity that I have had to help design and co-teach a large, online non-calculus based introductory statistics course for undergraduate students. Lastly, I will discuss how I conduct research on this course each semester, as well as preliminary results from the first two semesters of implementation.

EO054 Room Jessel MODERN ADVANCES IN CHANGE-POINT DETECTION Chair: Andreas Anastasiou

E0546: **Multiscale autoregression on adaptively detected timescales**
Presenter: **Piotr Fryzlewicz**, London School of Economics, United Kingdom
Co-authors: Rafal Baranowski

A multiscale approach to time series autoregression is proposed, in which linear regressors for the process in question include features of its own path that live on multiple timescales. We take these multiscale features to be the recent averages of the process over multiple timescales, whose number of spans are not known to the analyst and are estimated from the data via a change-point detection technique. The resulting construction, termed Adaptive Multiscale AutoRegression (AMAR) enables adaptive regularisation of linear autoregressions of large orders. The AMAR model permits the longest timescale to increase with the sample size, and is designed to offer simplicity and interpretability on the one hand, and modelling flexibility on the other. As a side result, we also provide an explicit bound on the tail probability of the L2 norm of the difference between the autoregressive coefficients and their OLS estimates in the AR(p) model with i.i.d. Gaussian noise when the order p potentially diverges with, and the autoregressive coefficients potentially depend on, the sample size. The R package amar provides an efficient implementation of the AMAR modelling, estimation and forecasting framework.

E0706: **A new approach for open-end sequential change point monitoring**
Presenter: **Tobias Kley**, University of Bristol, United Kingdom
Co-authors: Josua Goesmann, Holger Dette

A new sequential monitoring scheme for changes in the parameters of a multivariate time series is proposed. In contrast to procedures proposed in the literature which compare an estimator from the training sample with an estimator calculated from the remaining data, we suggest to divide the sample at each time point after the training sample. Estimators from the sample before and after all separation points are then continuously compared calculating a maximum of norms of their differences. For open-end scenarios our approach yields an asymptotic level α procedure, which is consistent under the alternative of a change in the parameter.

E0886: **Efficiency of rank-based change-point tests for long-range dependent time series**
Presenter: **Annika Betken**, Ruhr-Universitat Bochum, Germany
Co-authors: Martin Wendler

Change-point tests based on rank statistics are considered to test for structural changes in long-range dependent observations. Under the hypothesis of stationary time series as well as under the assumption of a structural change in the data, i.e. under (local) alternatives approaching the null hypothesis of no change, the asymptotic distributions of the corresponding test statistics are derived. These theoretical results are based on a uniform reduction principle for the sequential empirical process in a two-parameter Skorohod space equipped with a weighted supremum norm. Moreover, we compare the efficiency of rank tests resulting from the consideration of different score functions. Theoretical results are accompanied by simulation studies.

E1225: **Monitoring in one and two sample multivariate situations and some modifications**
Presenter: **Marie Huskova**, Charles University, Czech Republic
Co-authors: Zdenek Hlavka, Simos Meintanis

One and two sample multivariate change point detection procedures are considered. The proposed method is a L2-type criteria based on empirical characteristic functions. The focus is on on-line procedures. Asymptotic properties are presented together with results of a simulation study. The new method is also applied on a real data-set from the financial sector over a time period. Possible extensions will be discussed.

EO520 Room MAL 152 STATISTICAL METHODS FOR NEW-AGE INFERENCE PROBLEMS Chair: Gourab Mukherjee

E0419: **Univariate total variation denoising, trend filtering and multivariate Hardy-Krause variation denoising**
Presenter: **Adityanand Guntuboyina**, University of California, Berkeley, United States

Total variation denoising (TVD) is a popular technique for nonparametric function estimation. We will first present a theoretical optimality result for univariate TVD for estimating piecewise constant functions. We will then present related results for various extensions of univariate TVD including adaptive risk bounds for higher-order TVD (also known as trend filtering) as well as a multivariate extension via the Hardy-Krause Variation which avoids the curse of dimensionality to some extent. We will also mention connections to shape restricted function estimation.

E0728: **Second order Stein: SURE for SURE and other applications in high-dimensional inference**
Presenter: **Pierre Bellec**, Rutgers, United States

Stein's formula states that a random variable of the form $z^T f(z) div f(z)$ is mean-zero for all functions f with integrable gradient, where $div f$ is the divergence of the function f and z is a standard normal vector. A Second Order Stein formula is proposed to characterize the variance of such random variables. In the Gaussian sequence model, a remarkable consequence of Stein's formula is Stein's Unbiased Risk Estimate (SURE) of the mean square risk of almost any given estimator for the unknown mean vector. A first application of the Second Order Stein formula is an Unbiased Risk Estimate of the risk of SURE itself (SURE for SURE): a simple unbiased estimate provides information about the squared distance between SURE and the squared estimation error. SURE for SURE has a simple form and can be computed explicitly for differentiable estimator, for example the Lasso and the Elastic Net. Other applications of the Second Order Stein formula are provided in high-dimensional regression. This includes novel bounds on the variance of the size of the model selected by the Lasso, and a general semi-parametric scheme to de-bias an almost differentiable initial estimator in order to estimate a low-dimensional projection of the unknown regression coefficient vector.

E1333: **Assessment of the corroboration of an elaborate theory of a causal hypothesis using conjunctions of evidence factors**
Presenter: **Bikram Karmakar**, University of Florida, United States
Co-authors: Dylan Small

An elaborate theory of predictions of a causal hypothesis consists of several falsifiable statements derived from the causal hypothesis. Statistical tests for the various pieces of the elaborate theory help to clarify how much the causal hypothesis is corroborated. In practice, the degree of corroboration of the causal hypothesis has been assessed by a verbal description of which of the several tests provides evidence for which of the several predictions. This verbal approach can miss quantitative patterns. We develop a quantitative approach. We first decompose these various tests of the predictions into independent factors with different sources of potential biases. Support for the causal hypothesis is enhanced when many of these evidence factors support the predictions. A sensitivity analysis is used to assess the potential bias that could make the finding of the tests spurious. Along with this multi-parameter sensitivity analysis, we consider the partial conjunctions of the tests. These partial conjunctions quantify the evidence supporting various fractions of the collection of predictions. A partial conjunction test involves combining tests of the components in the partial conjunction. We find the asymptotically optimal combination of tests in the context of a sensitivity analysis. Our analysis of an elaborate theory of a causal hypothesis controls for the familywise error rate.

E0761: **Statistical considerations in Mendelian randomization**
Presenter: **Qingyuan Zhao**, University of Cambridge, United Kingdom

Mendelian randomization (MR) can give unbiased estimate of a confounded causal effect by using genetic variants as instrumental variables. The summary-data MR design is rapidly gaining popularity in practice due to the increasing availability of large-scale genome-wide association studies. We will present a comprehensive statistical approach to overcome the challenges. Motivated by exploratory data analysis, the summary-data MR problem will be formulated as a linear errors-in-variables regression with over-dispersion and occasional outliers. This means that none of the genetic instruments is valid in the strict sense. We will develop a class of statistical estimators with increased robustness to invalid instruments and maximal efficiency using weak instruments across the entire genome. We will further demonstrate visualization tools to detect meaningful effect heterogeneity. The new methods will be used to re-analyze several cardiometabolic diseases and risk factors, yielding new insights into the role of HDL particles (the "good" cholesterol) in coronary artery disease.

EO578 Room MAL 254 COMPUTATIONAL METHODS APPLIED TO THE ENVIRONMENT Chair: Peter Craigmile

E0739: **Spatio-temporal modelling of respiratory disease risk with changing spatial boundaries**
Presenter: **Eilidh Jack**, University of Glasgow, United Kingdom
Co-authors: Duncan Lee, Nema Dean

Spatio-temporal patterns in population-level disease risk are often estimated from data relating to a set of irregularly shaped areal units, such as electoral wards or census tracts. One shortcoming of traditional areal unit modelling techniques is that the areal units are themselves artificial units of spatial recording and can influence the spatial pattern observed in the data. That is, if the areal units changed then so would the results. This is known as the modifiable areal unit problem (MAUP). Another common problem in areal unit data of this type is that often there are changes to boundaries that occur during the time period for which data are available. Statistically, this poses a challenge since using data from before and after this change would lead to non-comparable inference due to spatial misalignment. A statistical framework is proposed for solving these problems, by using the areal unit data to obtain inference on the spatio-temporal pattern in disease risk on a regular grid. This framework is illustrated with a study on the spatio-temporal trends in health inequalities in respiratory disease risk in Glasgow, Scotland.

E0763: **Modeling sea level processes on the US Atlantic coast**
Presenter: **Candace Berrett**, Brigham Young University, United States
Co-authors: William Christensen, Steve Sain, Nathan Sandholtz, Davod Coats, Claudia Tebaldi, Hedibert Lopes

One of the major concerns engendered by a warming climate are changing sea levels and their lasting effects on coastal populations, infrastructures, and natural habitats. Sea levels are now monitored by satellites, but long term records are only available at discrete locations along the coasts. Sea levels and sea level processes must be better understood at the local level to best inform real-world adaptation decisions. We propose a statistical model that facilitates the inclusion of known sea level processes, such as sea level rise and seasonal cycles, and also accurately accounts for residual spatio-temporal processes, all together governing sea level behavior along the coast. By combining a spatially-varying coefficient modeling approach with spatio-temporal factor analysis methods in a Bayesian framework, the method represents the contribution of each of these processes and accounts for corresponding dependencies and uncertainties in a coherent way. Additionally, the model provides a consistent way to estimate these processes and sea level values at unmonitored locations along the coast. We show the outcome of the proposed model using thirty years of sea level data from thirty-eight stations along the Atlantic (east) coast of the United States. Among other results, our method estimates the rate of sea level rise to range from roughly 1 mm/year in the northern and southern regions of the coast to 5.4 mm/year in the middle region.

E1302: **Spatial point processes**
Presenter: **Janine Illian**, University of Glasgow, United Kingdom

In recent years point process methodology has become increasingly familiar to applied users. Spatial point processes have been originally developed within mathematical statistics as stochastic processes that have spatial point patterns as realisations. Nowadays, there has been a shift towards using them as a tool for modelling the locations of objects or events in space (and time) in practical scientific applications. This shift implies a change in the aims of the statistical analysis and in the focus of the associated statistical research. Within mathematical statistics, point processes form part of stochastic geometry and hence the aim is to define mathematically tractable processes that best mimic the geometric properties of a certain type of point pattern. In the context of applied statistics, however, the main aim of a modelling exercise is to answer scientific questions. Hence appropriate inference, model interpretation and model assessment are of primary concern. We explore this shift in focus, reviews recent progress in making point process methodology more relevant in practice and highlights opportunities for research. In the light of this, we discuss issues concerning model assumptions, model construction and model assessment and draw on a number of concrete examples from ecology and beyond for illustration.

E1312: **Statistical emulation to quantify uncertainties in tsunami modelling using high performance computing**
Presenter: **Serge Guillas**, University College London, United Kingdom

Solutions to the investigation of uncertainties in tsunami impacts in three settings are presented. First, we consider landslides as a source of tsunamis

from the Indus Canyon in the Western Indian Ocean. We employ statistical emulation, i.e. surrogate modelling, to efficiently quantify uncertainties associated with slump-generated tsunamis at the slopes of the canyon. We demonstrate that the emulator-based approach is an important tool for probabilistic hazard analysis since it can generate thousands of tsunami scenarios in few seconds, compared to days of computations on High Performance Computing facilities for a single run of the tsunami solver. We then examine future tsunami hazard from the Makran subduction zone in the Western Indian Ocean. We capture these phenomena in high resolution (down to 10m) using carefully constructed unstructured meshes for the port of Karachi. An emulator approximates the functional relationship between inputs and outputs maximum velocity and free surface elevation. Finally, we create emulators that respect the nature of time series outputs. We introduce here a novel statistical emulation of the input-output dependence of these computer models: functional registration and Functional Principal Components techniques improve the predictions of the emulator. We apply this approach to the high resolution tsunami wave propagation and coastal inundation for the Cascadia region in the Pacific Northwest.

EO510 Room Senate MODELLING EXTREMES Chair: Holger Rootzen

E1604: **The tail dependograph**
Presenter: **Cecile Mercadier**, Universita Lyon 1, France
Co-authors: Olivier Roustant
All characterizations of non-degenerate multivariate tail dependence structures are both functional and infinite-dimensional. Taking advantage of the Hoeffding-Sobol decomposition, we derive new indices to measure and summarize the strength of dependence in a multivariate extreme value analysis. The tail superset importance coefficients provide a pairwise ordering of the asymptotic dependence structure. We then define the tail dependograph, which visually ranks the extremal dependence between the components of the random vector of interest. For the purpose of inference, a rank-based statistic is derived and its asymptotic behavior is stated. These new concepts are illustrated with both theoretical models and real data, showing that our methodology performs well in practice.

E1686: **Practical left-tail correction for the GEV model**
Presenter: **Daniela Castro-Camilo**, University of Glasgow, United Kingdom
The generalized extreme value distribution (GEV) is a three parameters family that describes the asymptotic behaviour of properly normalized maxima of a sequence of independent and identically distributed random variables. If the shape parameter ξ is equal to zero, the GEV has infinite support, whereas if $\xi > 0$, the limiting distribution has a power-law decay with infinite upper endpoint but finite lower endpoint. In practical applications, we assume that the GEV is a reasonable approximation for the distribution of maxima over blocks and we fit it accordingly. This implies that GEV properties, such as finite lower endpoint in the case $\xi > 0$, are inherited by the original distribution of block maxima, which might not have bounded support. This issue is particularly problematic in the presence of covariates. We propose the blended GEV distribution with infinite support to tackle this usually overlooked issue. Using a Bayesian framework, we reparametrize the GEV to offer a more natural interpretation of the (possible covariate-dependent) model parameters. Independent priors over the new location and spread parameters produce a joint prior distribution for the original location and scale parameters, while a property-preserving penalized complexity prior approach is used for the shape parameter to avoid inconsistencies in the existence of first and second moments.

E1963: **Statistical inference for clusters of extremes: Disjoint vs. sliding blocks estimators**
Presenter: **Rafal Kulik**, University of Ottawa, Canada
Limit theorems for empirical cluster functionals are presented. We consider both disjoint and sliding block estimators. Conditions for consistency and asymptotic normality are provided in terms of tail and spectral tail processes and can be verified for a large class of multivariate time series, including geometrically ergodic Markov chains. Applications include asymptotic normality for the classical extremal index and recently introduced cluster indices. Results for multiplier bootstrap processes are also provided.

E1721: **Modelling panels of extremes**
Presenter: **Luca Trapin**, University of Bologna, Italy
Co-authors: Debbie Dupuis, Sebastian Engelke
Extreme value regression has been widely used over the last years to study the determinants of tail-risk events. Since such extreme events rarely occur, estimates of the model parameters are usually derived using a small number of observations, thus inducing high uncertainty. A class of panel regression models for the extremes is presented where the cross section of the data is pooled to obtain more reliable estimates of the regression coefficients. To account for possible unobserved heterogeneity in the data, we allow the extreme value panel regression model to have group-specific parameters, and design a likelihood-based clustering algorithm to recover the unknown group structure. A large simulation study assesses the finite sample properties of the panel maximum likelihood estimator and assesses the ability of the clustering algorithm to recover the unknown group structure. Finally, the usefulness of the new class of models is illustrated in several real-world examples.

EO342 Room CLO 203 NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE FOR DIRECTIONAL DATA Chair: Davy Paindaveine

E0711: **Adaptive optimal kernel density estimation for directional data**
Presenter: **Thanh Mai Pham Ngoc**, University Paris Sud Orsay, France
Nonparametric density estimation with directional data is considered. A new rule is proposed for bandwidth selection for kernel density estimation. The procedure is automatic, fully data-driven, and adaptive to the degree of smoothness of the density. An oracle inequality and optimal rates of convergence for the L2 error are derived. These theoretical results are illustrated with simulations.

E0413: **Projection-based uniformity tests for directional data**
Presenter: **Eduardo Garcia-Portugues**, Carlos III University of Madrid, Spain
Co-authors: Paula Navarro, Juan A Cuesta-Albertos
Testing uniformity of a sample supported on the hypersphere is one of the first steps when analyzing multivariate data for which only the directions (and not the magnitudes) are of interest. A projection-based class of uniformity tests on the hypersphere is introduced. The new class allows for extensions of circular-only uniformity tests and introduces the first instance of an Anderson-Darling test in the context of directional data. Asymptotics and optimality of the new class of tests are discussed. A simulation study corroborates the theoretical findings. Finally, a real data example illustrates the usage of the new tests.

E1161: **Bandwidth-free pilot pre-smoothing for circular data**
Presenter: **Jose E Chacon**, Universidad de Extremadura, Spain
Co-authors: Carlos Tenreiro
Least-squares cross validation (LSCV) yields a simple automatic bandwidth selection technique for kernel density estimation with circular data. Likelihood cross validation (LCV) represents another viable alternative, which in addition has shown a more efficient behaviour in practice. A smoothed cross validation (SCV) procedure is investigated, in which pilot pre-smoothing is employed to improve the performance of LSCV. However, contrary to the usual scenario, no pilot bandwidth or reference distribution is needed, so that the new method is fully data-driven. The practical performance of this SCV approach is compared with the traditional selectors in an extensive simulation study.

E1531: **Circular data with error-in-variables**
Presenter: **Marco Di Marzio**, University of Chieti-Pescara, Italy
Co-authors: Stefania Fensore, Agnese Panzera, Charles C Taylor
Nonparametric methods are discussed for the case when data are observed with error and have a circular nature. Some classical approaches are explored, such as the deconvolution one, but also less popular ones, like bias reduction under the hypotesis of double asymptotics, and finally, some new resampling strategies. Proposals are justified by both asymptotic properties and simulative evidences.

EO092 Room CLO 204 STATISTICS IN NEUROSCIENCE I Chair: Jeff Goldsmith

E0198: **Statistical approaches for disentangling the nature of brain lesions**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States
Lesions in the white matter of the brain, including those that arise in multiple sclerosis, are abnormalities measurable on MRI. While much work in the statistical literature has focused on the identification of these lesions, less work has focused on the nature of these lesions. As new imaging modalities arise that allow us to better interrogate these lesions, new statistical modeling problems that include spatial constraints and overlapping domains of analysis are increasingly important. Leveraging multi-modal imaging approaches that focus on knowledge about etiology is critical for developing the next generation of robust and generalizable imaging biomarkers.

E0657: **A grouped Beta process model for multivariate resting-state EEG microstate analysis on twins**
Presenter: **Mark Fiecas**, University of Minnesota, United States
EEG microstate analysis investigates the collection of distinct temporal blocks that characterize the electrical activity of the brain. We propose a Bayesian nonparametric model that estimates the number of microstates and the underlying behavior. We use a Markov switching vector autoregressive (VAR) framework, where a hidden Markov model (HMM) controls the non-random state switching dynamics of the EEG activity and a VAR model defines the behavior of all time points within a given state. We analyze resting state EEG data from twin pairs collected through the Minnesota Twin Family Study, consisting of 70 epochs per participant corresponding to 140 seconds of EEG data. We fit our model at the twin pair level, sharing information within epochs from the same participant and within epochs from the same twin pair. We capture within twin pair similarity by using a Beta process Bernoulli process to consider an infinite library of microstates and allowing each participant to select a finite number of states from this library. The state spaces of highly similar twins may completely overlap while dissimilar twins could select distinct state spaces. In this way, our Bayesian nonparametric model defines a sparse set of states which describe the EEG data. All epochs from a single participant use the same set of states and are assumed to adhere to the same state switching dynamics in the HMM model.

E0718: **Intensity warping for multisite MRI harmonization**
Presenter: **Julia Wrobel**, University of Colorado Denver, United States
Co-authors: Jeff Goldsmith, Russell Shinohara, Melissa Martin
In multisite neuroimaging studies there is often unwanted technical variation across scanners and sites. These scanner effects can hinder detection of biological features of interest, produce inconsistent results, and lead to spurious associations. We assess scanner effects in two brain magnetic resonance imaging (MRI) studies where subjects were measured on multiple scanners within a short time frame, so that one could assume any differences between images were due to technical rather than biological effects. We propose mica (multisite image harmonization by CDF alignment), a tool to harmonize images taken on different scanners by identifying and removing within-subject scanner effects. The goals were to (1) establish a method that removes scanner effects by leveraging multiple scans collected on the same subject, and, building on this, (2) develop a technique to quantify scanner effects in large multisite trials so these can be reduced as a preprocessing step. We found that unharmonized images were highly variable across site and scanner type, and our method effectively removed this variability by warping intensity distributions. We further studied the ability to predict intensity harmonization results for a scan taken on an existing subject at a new site using cross-validation.

E1222: **Aging multiple sclerosis lesions on structural magnetic resonance images**
Presenter: **Elizabeth Sweeney**, Weill Cornell, United States
Brain structural magnetic resonance imaging (sMRI) is a tool that uses a magnetic field to produce detailed images of the brain. Patients with multiple sclerosis (MS) have lesions in their brains which are visible on sMRI. Part of the 2017 McDonald diagnostic criteria for MS requires that a patient have one clinical MS attack as well as lesions that display dissemination in time and space. Patients would like to have a diagnosis as soon as possible, but showing dissemination in time is difficult from a single MRI scan and requires the invasive technique of administering intravenous gadolinium for an MRI scan. We have developed an algorithm for estimating the age of multiple sclerosis lesions using non-invasive sMRI and Quantitative Susceptibility Mapping (QSM) MRI from a single MRI study. The algorithm creates numerous radiomic features from the sMRI and QSM, performs principal component analysis (PCA) for dimension reduction, and then feeds into a random forest which predicts whether a lesion is older or younger than a year. In a validation cohort, the algorithm achieved an area under the receiver operating characteristic curve (AUC) of .90. The results show promise for revising the diagnostic criteria of MS and allowing for a faster and less invasive diagnosis.

EO282 Room MAL 251 HIGH DIMENSIONAL AND LATENT VARIABLE REGRESSION MODELING Chair: Valentin Todorov

E1305: **Instance-dependent cost-sensitive logistic model for detecting transfer fraud**
Presenter: **Tim Verdonck**, UAntwerp, Belgium
Co-authors: Sebastiaan Hoppner, Bart Baesens
Credit card fraud is a growing problem that affects card holders around the world. Financial institutions are, therefore, forced to continuously improve their fraud detection systems and they do so by increasingly relying on predictive models. The aim of detecting transfer fraud is to identify transactions with a high probability of being fraudulent. The task of predicting the fraudulent nature of transactions can be presented as a binary classification problem. Different solutions for detecting fraud are then commonly evaluated based on some sort of misclassification measure, and do not take into account the actual financial costs associated with the fraud detection process. Fraud detection, however, is a typical example of cost-sensitive classification, where the costs due to misclassification vary between instances. Nevertheless, current transfer fraud detection algorithms often miss including the real costs associated with credit card fraud. Based on an instance-dependent cost matrix for transfer fraud detection, a cost measure is introduced that represents the monetary gains and losses due to the classification of credit transfers. We present a classifier that minimizes this instance-dependent cost measure directly into the model construction during the training step, where the classifier's interior model structure resembles a lasso-regularized logistic regression. As an illustration, we compare our proposed method against existing credit card fraud detection models.

E1364: **A PLS method for seeking canonical correlations in case of perfect multicollinearity**
Presenter: **Michele Gallo**, University of Naples Orientale, Italy
Co-authors: Violetta Simonacci
Canonical correlation analysis (CCA) is a useful tool for investigating the relationships between two sets of variables. If dispersion matrices can be inverted, canonical variates with maximal correlation are generally identified by means of singular value decomposition. However, when one or both variable groups are compositional, this classical approach cannot be followed. Compositional data are positive values which carry relative information describing the parts of a whole. In consequence they present a perfectly multicollinear structure and are characterized by singular dispersion matrices. As a solution to this issue which excludes a standard approach, an alternative way of computing canonical variates is

proposed. Data are first transformed in log-ratio coordinates, then the Partial Least Squares approach is applied. This method provides a fast and easy way to deal with non-invertible dispersion matrices and, in addition, it yields results which are easy to interpret. The proposed methodology is assessed in an experimental study in which a comparison among alternative PLS algorithms is also provided, namely NIPALS, SIMPLS and Kernel.

E1449: **Weighting of logratios in compositional regression**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Nikola Stefelova, Javier Palarea-Albaladejo

Compositional data are multivariate positive observations which are characterized by the scale invariance property: any positive multiple does not change the essential information contained in ratios between components. As such, compositional data represent observations carrying relative information, commonly expressed in units like proportions, percentages, mg/kg, mg/l and so on; thus including also those which do not necessarily impose a constant sum constraint of components. The logratio analysis approach to compositional data builds on a Euclidean space structure for scale invariant observations, so-called the Aitchison geometry, and expresses compositions in real coordinates, preferably with respect to an orthonormal basis. One particular choice are pivot coordinates, by which the first coordinate aggregates all logratios with respect to a component of interest. Because some logratios may be affected by data quality problems, or may represent a completely different process than the others, it is desirable in practice to be able to reduce their contribution to the first pivot coordinate. Therefore, in a context of regression analysis with compositional explanatory variables, weights are sensibly defined according to correlations between a (real) response and the logratios. Theoretical aspects will be accompanied by demonstrations with both simulated and real-world metabolomic data.

E1915: **Bayesian topic regression: An econometric model for inference with heterogeneous high dimensional data**
Presenter: **Julian Ashwin**, University of Oxford, United Kingdom
Co-authors: Maximillian Ahrens

When incorporating text data into econometric models, a fundamental question is how to represent and select text features. Dictionaries and other unsupervised feature selection approaches have been widely used in financial and economic modelling to provide interpretable measures of interest from text data. However these methods are often not optimised for the research question at hand, since feature selection is performed separately from subsequent econometric analysis, thereby potentially discarding information relevant to the inference task. We combine a supervised LDA topic model with a multivariate Bayesian regression framework, allowing us to simultaneously perform feature extraction and parameter estimation. This has several advantages over existing supervised feature selection methods. First, our Bayesian approach allows inference on the coefficients of the regression model which takes into account the sampling uncertainty involved in the text feature estimation. Second, by estimating coefficients on text and non-text covariates jointly, it respects the Frisch-Waugh-Lovell Theorem which prevents simply using a residualised dependent variable. Third, we are able to use information from observations without text documents. Finally, our model allows meaningful inference in situations of a very high dimensional feature space, but a relatively small number of observations. We demonstrate this on synthetic data and on central bank communication data.

EO604 Room MAL 252 ADVANCES IN PRECISION AND COVARIANCE MATRIX ESTIMATION Chair: Eugen Pircalabelu

E0807: **Cellwise robust regularized precision matrices for discriminant analysis**
Presenter: **Ines Wilms**, Maastricht University, Netherlands
Co-authors: Stephanie Aerts

Quadratic and Linear Discriminant Analysis (QDA/LDA) are the most often applied classification rules under normality. In QDA, a separate covariance matrix is estimated for each group. If there are more variables than observations in the groups, the usual estimates are singular and cannot be used anymore. Assuming homoscedasticity, as in LDA, reduces the number of parameters to estimate. This rather strong assumption is however rarely verified in practice. Regularized discriminant techniques that are computable in high-dimension and cover the path between the two extremes QDA and LDA have been proposed in the literature. However, these procedures rely on sample covariance matrices. As such, they become inappropriate in presence of cellwise outliers, a type of outliers that is very likely to occur in high-dimensional datasets. We propose cellwise robust counterparts of these regularized discriminant techniques by inserting cellwise robust covariance matrices. Our methodology results in a family of discriminant methods that (i) are robust against outlying cells, (ii) provide, as a by-product, a way to detect outliers, (iii) cover the path between LDA and QDA, and (iv) are computable in high-dimensions.

E0851: **The normal scores estimator for the high-dimensional Gaussian copula model**
Presenter: **Yue Zhao**, KU Leuven, Belgium
Co-authors: Christian Genest

The (semiparametric) Gaussian copula model consists of distributions that have dependence structure described by Gaussian copulas but that have arbitrary marginals. A Gaussian copula is in turn determined by a Euclidean parameter R called the copula correlation matrix. We study the normal scores (rank correlation) estimator, also known as the van der Waerden estimator, of R in high dimensions. It is well known that in fixed dimensions, the normal scores estimator is the optimal estimator of R, i.e., it has the smallest asymptotic covariance. Curiously though, in high dimensions, nowadays the preferred estimators of R are usually based on Kendall's tau or Spearman's rho. We show that the normal scores estimator in fact remains the optimal estimator of R in high dimensions. More specifically, we show that the approximate linearity of the normal scores estimator in the efficient influence function, which in fixed dimensions implies the optimality of this estimator, holds in high dimensions as well.

E0937: **Community detection on precision matrices with group-based penalties**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium
Co-authors: Gerda Claeskens

A new strategy for probabilistic graphical modeling is developed that draws parallels from social network analysis. Probabilistic graphical modeling summarizes the information coming from multivariate data in a graphical format where nodes, corresponding to random variables, are linked by edges that indicate dependence relations between the nodes. The purpose is to estimate the structure of the graph (which nodes connect to which other nodes) when data at the nodes are available. On the opposite side of the spectrum, social network analysis considers the graph as the observed data. Given thus the graph where connections between nodes are observed rather than estimated, social network analysis estimates models that represent well an underlying mechanism which has generated the observed graph. We propose a new method that exploits the strong points of each framework as it estimates jointly an undirected graph, based on a precision matrix, and communities of homogenous nodes. The structure of the communities is taken into account when estimating the precision matrix and, conversely, the structure of the graph is accounted for when estimating homogeneous communities of nodes. The procedure uses a joint group graphical lasso approach with community detection-based grouping, such that some groups of edges co-occur in the estimated graph. The grouping structure is unknown and is estimated based on community detection algorithms.

E1040: **RSVP-graphs: Fast high-dimensional covariance matrix estimation under latent confounding**
Presenter: **Rajen D Shah**, University of Cambridge, United Kingdom
Co-authors: Benjamin Frot, Gian-Andrea Thanei, Nicolai Meinshausen

The focus in on the problem of estimating a high-dimensional $p \times p$ covariance matrix Σ, given n observations of confounded data with covariance

$\Sigma + \Gamma\Gamma^T$, where Γ is an unknown $p \times q$ matrix of latent factor loadings. We propose a simple and scalable estimator based on the projection on to the right singular vectors of the observed data matrix, which we call RSVP. Our theoretical analysis of this method reveals that in contrast to approaches based on removal of principal components, RSVP is able to cope well with settings where the smallest eigenvalue of $\Gamma^T\Gamma$ is relatively close to the largest eigenvalue of Σ, as well as when eigenvalues of $\Gamma^T\Gamma$ are diverging fast. RSVP does not require knowledge or estimation of the number of latent factors q, but only recovers Σ up to an unknown positive scale factor. We argue this suffices in many applications, for example if an estimate of the correlation matrix is desired. We also show that by using subsampling, we can further improve the performance of the method. We demonstrate the favourable performance of RSVP through simulation experiments and an analysis of gene expression datasets collated by the GTEX consortium.

EO348 Room MAL 253 SOCIETAL IMPLICATIONS OF WORK IN STATISTICS AND DATA SCIENCE Chair: Jennifer Hill

E1200: **Bayesian survey design: A new paradigm for social science research**
Presenter: **Seth Flaxman**, Imperial College London, United Kingdom

Some preliminary work on a Bayesian active learning approach to survey design is described. In many social science fields, the existing gold standard for surveys is a stratified random sampling (i.e. some form of Monte Carlo sampling). Once the data has been collected, regression is used to model a response surface of interest. Inspired by the field of Bayesian experimental design and by recent advances in the machine learning literature on Bayesian Optimization and Bayesian Quadrature, we propose a more efficient alternative to Monte Carlo: "Bayesian survey design". Using prior information (e.g. from previous surveys or initial rounds of an ongoing survey), we model the unknown response surface using a Bayesian model (e.g. GLM, Gaussian processes, or Bayesian Additive Regression Trees). Then, conditional on this model, we take an active learning approach to select future respondents so as to optimize some objective function, e.g. we minimize the posterior variance. We present simulation results to demonstrate our approach and preliminary results on a real dataset.

E1514: **A framework to analyze and dissect dissemination of motivated information through social media**
Presenter: **Sunandan Chakraborty**, Indiana University IUPUI, United States

'Fake news' has been an important part of our contemporary public discourse. The term has been used by a variety of actors to broadcast items that mislead people in the guise of legitimacy. In India, with a large share of new social media users, information that is either blatantly false, or motivated to inflame, can spread to a large population within a small span of time. Consuming fake news has led to serious consequences, including deaths in India. On many platforms, such as, WhatsApp, due to encryption of messages, it becomes extremely difficult for law enforcement to intervene and stop such information from spreading. At present, most of this detection is carried out manually. With the increasingly fast rate of generation and even faster rate of spread of such stories, it is infeasible to rely purely on manual interventions to address this problem. The first focus will be on the nature of false information that is spread in India by analyzing a collection of posts verified to be false. The methods and intent used in the spread of false information will be presented. The second focus will be on methods to collect data from secure platforms like Whatsapp using crowdsourcing and develop methods to detect false information within shared posts.

E1559: **Evaluating predictive bias in the presence of differential label noise**
Presenter: **Alexandra Chouldechova**, Carnegie Mellon University, United States

Risk assessment tools are widely used around the country to inform decision making within the criminal justice system. Recently, considerable attention has been devoted to the question of whether such tools may suffer from racial bias. In this type of assessment, a fundamental issue is that the training and evaluation of the model is based on a variable (arrest) that may represent a noisy version of an unobserved outcome of more central interest (offense). We present a sensitivity analysis framework for assessing how differential label noise affects the predictive bias properties of the risk assessment model as a predictor of reoffense. We also discuss the impact of differential label noise on the model training process.

E1560: **The effect of London cycle superhighways on traffic congestion**
Presenter: **Emma McCoy**, Imperial College London, United Kingdom

The Mayor of London's Transport Strategy set out the aim for cycling journeys in London to increase from 2% of all journeys in 2001 to 5% by 2026. Various interventions have been introduced in an attempt to meet this target, including the introduction of cycle superhighways. In addition to encouraging cycling it has been claimed that cycle superhighways will reduce traffic congestion. Estimating the effect of the cycle superhighways on congestion is complicated due to the non-random treatment assignment over the transport network. We outline a propensity score based method for the analysis of pre- and post-intervention data and demonstrate through simulations studies the superiority of the proposed method over existing competitors. The method is applied to analyse pre and post-intervention data collected on the London transport network which includes traffic and road characteristics along with socio-economic factors.

EO056 Room SH349 SPORT ANALYTICS Chair: Andreas Groll

E0250: **Evaluating sports tournament predictions**
Presenter: **Claus Ekstrom**, University of Copenhagen, Denmark

Predicting the winner of a sports tournament has become an ever-increasing challenge for researchers, sports fans and the growing business of bookmakers. Before the start of major tournaments, such as the FIFA World Cup, the Australian Open, or the IHF Handball Championship, the world press is discussing the various predictions. The quality of the predictions is evaluated after the tournaments. A novel way to evaluate the quality and precision of pre-tournament predictions is presented. The best prediction is the one that most closely resembles the actual outcome of the tournament while still avoiding making confident - but wrong - claims about the outcome. This new tournament rank prediction score (TRPS) will be illustrated by comparing different predictions from the 2018 FIFA World Cup. We will discuss how the TRPS can be used to optimally combine various predictions methods by assigning weights to each prediction in such a way that the TRPS becomes maximal.

E0289: **Analyzing positional data from soccer games with deep reinforcement learning**
Presenter: **Ulf Brefeld**, Leuphana University of Luneburg, Germany

Positional data are regularly recorded in many soccer leagues, such as Bundesliga. The data are captured at 25fps and contains (x,y) coordinates of all players and the ball. Although that data are available for some years, they are mainly used for computing heat maps showing the whereabouts of players or descriptive statistics like the covered kilometers by a player. By contrast, we will show that the data can actually be used for more sophisticated analyses using state-of-the-art deep architectures. We will present examples dealing with player movement models, controlled zones, and predictions of player actions.

E0386: **A hybrid random forest to predict soccer matches in international tournaments**
Presenter: **Hans Van Eetvelde**, Ghent University, Belgium
Co-authors: Andreas Groll, Christophe Ley, Gunther Schauberger

A new hybrid modeling approach is proposed for the scores of international soccer matches which combines random forests with Poisson ranking methods. While the random forest is based on the competing teams covariate information, the latter method estimates ability parameters on historical match data that adequately reflect the current strength of the teams. We compare the new hybrid random forest model to its separate building blocks as well as to conventional Poisson regression models with regard to their predictive performance on all matches from the four FIFA World Cups 2002-2014. It turns out that by combining the random forest with the team ability parameters from the ranking methods as an

additional covariate the predictive power can be improved substantially. Finally, the hybrid random forest is used (in advance of the tournament) to predict the FIFA World Cup 2018. To complete our analysis on the previous World Cup data, the corresponding 64 matches serve as an independent validation data set and we are able to confirm the compelling predictive potential of the hybrid random forest which clearly outperforms all other methods including the betting odds.

E0833: **A regularized hidden Markov model for analysing the "hot shoe" in football**
Presenter: **Marius Oetting**, Bielefeld University, Germany
Co-authors: Andreas Groll

Although academic research on the "hot hand" effect (in particular, in sports, especially in basketball) has been going on for more than 30 years, it still remains a central question in different areas of research whether such an effect exists. We investigate the potential occurrence of a "hot shoe" effect for the performance of penalty takers in football based on data from the German Bundesliga. For this purpose, we consider hidden Markov models (HMMs) to model the (latent) forms of players. To further account for individual heterogeneity of the penalty taker as well as the opponent's goalkeeper, player-specific abilities are incorporated in the model formulation together with a LASSO penalty. The results suggest states which can be tied to different forms of players, thus providing evidence for the hot shoe effect, and shed some light on exceptionally well-performing goalkeepers, which are of potential interest to managers and sports fans.

EC814 Room G21A CONTRIBUTIONS IN BIOSTATISTICS Chair: Mireille Schnitzer

E1250: **Mediation analysis when outcome and mediator are semi-competing events with application in health disparities research**
Presenter: **Linda Valeri**, Columbia University, United States
Co-authors: Cecile Proust-Lima, Helene Jacqmin-Gadda

Novel methodology for mediation analysis is proposed to explain how much of the effect of the exposure on a terminal time-to-event outcome, say death, is attributed to the non-terminal potential intermediate time-to-event. Addressing this question is important in health disparities research when we seek to quantify inequities in access to high quality and timely treatment and their impact on patients survival time. We formalize a type of direct and indirect effects using the potential outcome framework in the presence of semi-competing risks. Mediation is studied in a multistate model in continuous time. Monte Carlo simulation based as well as closed form estimators of the causal contrasts are developed. We show via simulations that mediation analysis ignoring censoring in mediator and outcome time-to event-processes and/or ignoring competing risks may give misleading results. Rigorous definition of the direct and indirect effects and estimation of the joint outcome and mediator distributions in the presence of semi-competing risks is crucial for valid investigation of mechanisms in continuous time. We employ this novel methodology to investigate the role of delaying treatment uptake in explaining racial disparities in cancer survival in a multi-center cohort study of colorectal cancer patients.

E1621: **Interval estimates of event probability from pairwise correlated data: Application in epidemiology of birth defects**
Presenter: **Jan Klaschka**, Institute of Computer Science of the Czech Academy of Sciences, Czech Republic
Co-authors: Marek Maly, Antonin Sipek

Interval estimates of event probability are studied within a generalization of Bernoulli trials model: $n = 2m$ zero-one-valued variables consist of m pairs with correlation ϕ between the two components. Independence between the m pairs and a common expectation θ of all n variables are assumed. The primary motivation and the main application field is in the epidemiology of congenital anomalies (birth defects) in twins. Occurrence of birth defects in both twins is known to be more frequent than under independence. Ignoring the fact and applying the binomial model would lead to over-liberal inferences. The focus is on the computation of exact interval estimates of θ - so far for fixed ϕ. Numerical procedures have been designed for the calculation of confidence bounds of Clopper-Pearson, Sterne and Blaker types. The key building block is the calculation of the probability mass function (pmf) of the number of events. Several pmf calculation methods have been tested. Among them, the numerical inversion (via iFFT) of the characteristic function appears to be the most computationally effective. A quasi-symbolic calculation based on the pmf representation as a matrix of polynomial coefficients is competitive under some (but not all) settings.

E1723: **Variable selection method for the logistic regression model using a model-X knockoffs algorithm**
Presenter: **Takafumi Nakatsu**, Chuo University, Japan
Co-authors: Toshinari Kamakura

In the field of medical research, the logistic regression model is commonly used for binary response variables, such as, treatment success or failure. It is important to select covariates which have substantial relation to the response variables to use the model. Covariates are usually selected with the standard likelihood-based testing methods, the likelihood ratio test, score test, and Wald test, by exploring a combination of covariates which gives true non-zero regression parameters. However, until now, the optimality of selected covariates has not yet been well investigated. Recently, a novel statistical procedure, model-X knockoffs algorithm has provided a way to identify important covariates under controlling False Discovery Rate (FDR). We developed a new method to reach an optimum solution of covariate combination with true non-zero regression parameters, employing the model-X knockoffs algorithm. Features of the method is being investigated in comparative simulation studies. The results of the studies will be presented and discussed.

E1909: **Multivariate and multiscale complexity of long-range correlated cardiovascular variability signals**
Presenter: **Ana Paula Rocha**, Univ Porto - Fac Ciencias and CMUP, Portugal
Co-authors: Celestino Amado, Aurora Martins, Maria Eduarda Silva

An intrinsic feature of most physiological systems, as well of some climatic or econometric systems, is their dynamical complexity, resulting from the combined activity of several coupled mechanisms typically operating across multiple temporal scales. The cardiovascular system is one of such systems and specific complex characteristics such as long memory have been considered from a model based autoregressive fractionally integrated (ARFI) parametric approach. Entropy rate is another current measure of complexity. Recently, a computationally reliable approach for the practical calculation of the linear multiscale entropy (MSE) of cardiovascular variability signals was introduced. This approach explores a state space formulation and is also able to attend the simultaneous presence of short-term dynamics and long-range correlations by using the ARFI modeling. Moreover, given the interactions present in these systems, it is expectable that a multivariate approach provides enhanced descriptions and a natural generalization considers a multivariate approach with vector ARFI models (VARFI). The methods are applied in some typical experimental/clinical stress situations using cardiovascular signals, and the new measures appear to reflect the changes in the cardiovascular variability system dynamics.

EC810 Room MAL 355 CONTRIBUTIONS IN STATISTICAL LEARNING METHODS AND APPLICATIONS Chair: Charles C Taylor

E1926: **Resampling methods for comparing clustering solutions**
Presenter: **Pietro Coretto**, Universita di Salerno, Italy
Co-authors: Luca Coraggio

Selecting an optimal clustering solution is a longstanding problem. In model-based clustering this amounts to choose the architecture of the model mixture distribution. The main decisions to be made are: the cluster prototype distribution, the number of mixture components, and often other restrictions on the clusters' geometry. Classical penalized model selection criteria based on the observed likelihood function have been proposed to address this issue. We propose a selection strategy based on resampling, and we compare it with classical methods.

E1830: **The evaluation of statistical learning models in macroeconomic time series**
Presenter: **Simone Tonini**, Sant Anna School of Advanced Studies - Pisa, Italy

Thanks to the great amount of available data a part of the econometric literature focused on the application of statistical learning models in macroeconomics forecasting. Despite the interesting results obtained on forecasting accuracy, few works check whether these models are also able to get a good representation of the true data generating process. This kind of analysis is not trivial, specially when few variables determine the outcome in contexts with multicollinearity and serially correlated data. Moreover, understanding the role of each variable on the data generating process is crucial when the purpose is to understand and replicate the evolution and the dynamics of certain complex economic phenomena. To clarify this point, a large simulation study is presented in order to compare the performances in forecasting, variable selection and coefficient estimation accuracy of a large set of models under several data generating process and signal-to-noise ratio. Furthermore, the role of the serial correlation in the formation of spurious correlations is studied and a serial-decorrelation pre-step is proposed, in order to evaluate possible improvements in variable selection and forecasting accuracy using the i.i.d. component of the time series data.

E1793: **Learning multiple quantiles with neural networks**
Presenter: **Sang.Jun Moon**, University of Seoul, Korea, South
Co-authors: Jong-june Jeon, Jason Sang Hun Lee, Yongdai Kim

A neural network model is presented to estimate multiple conditional quantiles satisfying the non-crossing property. Motivated by linear non-crossing quantile regression, we apply inequality constraints used in the developed model to learning the neural network with a feasible set. In particular, to use the first-order optimization method on the feasible set, we develop a modified version of the interior point method. In the algorithm, an auxiliary variable lying on the feasible set is introduced as a proxy of the model parameter in the barrier function. By regularizing the difference between the auxiliary and original parameters, our proposed algorithm achieves close to the optimal solution while avoiding the projected gradient step with polynomial computation to significantly improve the computational efficiency. We compare predictive performances of multiple quantiles regression with neural networks with those of existing neural network models and apply our proposed model to the prediction of real data.

E1928: **Predicting shape of dives of Southern elephant seals using functional regression tree models**
Presenter: **Morgan Godard**, MIO-Mediterranean Institute of Oceanography, France
Co-authors: Claude Mante, Christophe Guinet, David Nerini

In recent years, the study of animal movements in the ocean has been revolutionized with massive use of miniature measuring devices providing access to complex behavioral data and associated environmental data sampled at very high frequency. These data are called functional data. The objectives are to highlight the relationships between dives of Southern elephant seals, Mirounga leonina, and the physical environment in which elephant seals operate. Starting from a huge data set of elephant seal dives, we first show how to construct functional dive profiles from point-wise samples. We then propose a generalized regression tree method where the predictive variable is a function. Regression tree models are built to predict the shape of dives using discrete environmental variables (i.e. temperature at a depth of 250m) and environmental profiles (i.e. temperature, salinity) as predictors. The connection between shapes of dives and shapes of environmental profiles will be highlighted. Tree capabilities for predictor selection and the choice of splitting criterion will be discussed.

EG571 Room MAL 354 CONTRIBUTIONS IN CLUSTERING AND CLASSIFICATION Chair: Christian Hennig

E1595: **A new split criterion for classification trees with binary data**
Presenter: **Abdulmajeed Alharbi**, Durham University, United Kingdom
Co-authors: Frank Coolen, Tahani Coolen-Maturi

Classification is a technique that is used to assign an observation to one of a set of predefined categories. Classification trees are considered to be one of the most popular approaches for classification. A new classification method is introduced, namely, the Direct Nonparametric Predictive Inference (D-NPI) classification for binary data. Nonparametric Predictive Inference (NPI) is a statistical method which uses few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty. D-NPI is based on the NPI lower and upper probabilities without adding any further assumptions or information. A new procedure for building classification trees using the NPI method is presented. It uses a new split criterion, called correct indication, for constructing classification trees. Lower and upper probabilities of correct indication are provided using the NPI method for Bernoulli data. The aim is to maximize the probability of correct indication for a future observation. Imprecision, the difference between upper and lower probabilities is used as stopping criterion. An experiment is carried out to compare this new procedure with classical classification trees. Initial comparisons with alternative methods suggest that the D-NPI classification performs well and tends to lead to relatively small trees.

E1598: **Classification trees with NPI-based thresholds**
Presenter: **Masad Alrasheedi**, Durham university, United Kingdom
Co-authors: Tahani Coolen-Maturi, Frank Coolen

Nonparametric Predictive Inference (NPI) is a statistical method which uses few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty. Applications of NPI for building classification trees where the inference itself is based on future observations are presented. When building classification trees, it is necessary to handle continuous attributes and select the optimal thresholds for each continuous variable. One technique that is commonly used to find the optimal thresholds is the C4.5 algorithm. This algorithm suggests sorting the data set and calculates the entropy for all midpoints between consecutive values. At each level, C4.5 chooses the attribute that maximizes entropy. However, classical methods usually do not focus on future observations, only on the data at hand where the attribute value of observation is known. A new method for selecting the optimal thresholds by using the NPI approach is presented. Moreover, the classical approaches often choose the split variables by maximising expected entropy. A new technique is introduced by using the NPI approach. In this technique, the full range of expected entropy is taken for each variable. Initial comparisons of the new approach with alternative methods indicate good classification performance, and the resulting trees are relatively small.

E1766: **Non-metric unfolding on augmented data matrix: A copula-based approach**
Presenter: **Marta Nai Ruscone**, LIUC, Italy
Co-authors: Antonio Dambrosio

Unfolding applies multidimensional scaling to an off-diagonal *nxm* matrix, representing the scores (or the rank) assigned to a set of *m* items by *n* individuals or judges. The goal is to obtain two configurations of points representing the position of the judges and the items in a reduced geometrical space. Each point, representing each individual, is considered as an ideal point so that its distances to the object points correspond to the preference scores. Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions. To tackle the problem, several methods have been proposed. By following a previous approach, we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms. In order to augment the data matrix, we use copulas-based association. Both experimental evaluations and applications to well-known real data sets show that the proposed strategy produces non-degenerate non-metric unfolding solutions.

E1893: **A customer segmentation method for UK open banking-type data**
Presenter: **Andrej Svetlosak**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Miguel de Carvalho
An unsupervised method is developed to segment individuals based on demographics, financial status and financial behaviour tailored for transactional open banking-type data. We analyse transactional records of 9,875 customers in 2017 and their registered accounts across 49 UK financial institutions. We propose a segmentation method that takes advantage of the information hierarchy between demographics, financial status and financial behaviour, creating a three-level segmentation that is easy to interpret. After obtaining the segments we assess their quality and highlight how the groups differ in demographics, financial status and financial behaviour.

EP002 Room Macmillan Hall and Crush Hall POSTER SESSION CMSTATISTICS I Chair: Elena Fernandez Iglesias

E1546: **Improved Monte Carlo inference methods for network meta-analysis**
Presenter: **Hisashi Noma**, The Institute of Statistical Mathematics, Japan
Co-authors: Kengo Nagashima
Network meta-analysis enables comprehensive synthesis of evidence concerning multiple treatments and their simultaneous comparisons based on both direct and indirect evidence. In the practice of network meta-analyses, multivariate random effects models have been routinely used for addressing between-studies heterogeneities. Although their standard inference methods depend on large sample approximations (e.g., restricted maximum likelihood [REML] estimation) for the number of trials synthesized, the numbers of trials are often moderate. In these situations, standard estimators cannot be expected to behave in accordance with asymptotic theory; in particular, confidence intervals cannot be assumed to exhibit their nominal coverage probabilities (also, the type-I error probabilities of the corresponding tests cannot be retained). The invalidity issue may seriously influence the overall conclusions of network meta-analyses. We provide permutation-based Monte Carlo inference methods that enable exact joint inferences for average outcome measures without large sample approximations. We also provide accurate marginal inference methods under general settings of network meta-analyses. We propose effective approaches for permutation inferences using optimal weighting based on the efficient score statistic. The effectiveness of the proposed methods is illustrated via applications to a network meta-analysis for antihypertensive drugs on incident diabetes.

E1636: **On a new goodness-of-fit test for the Rayleigh distribution based on a conditional expectation characterization**
Presenter: **James Allison**, Northwest University, South Africa
Co-authors: Joseph Ngatchou-Wandji, Shawn Liebenberg
New goodness-of-fit tests for the Rayleigh distribution are proposed which are based on a characterization involving a conditional expectation. The asymptotic properties of the test are explored, and the performance of the new test is evaluated and compared to that of existing tests by means of a Monte Carlo study. It is found that the newly proposed tests perform satisfactory compared to the competitor tests.

E1651: **Asymptotic behaviour of auto- and cross-dependence measures for heavy-tailed time series**
Presenter: **Aleksandra Grzesiek**, Wroclaw Univeristy of Science and Technology, Poland
Co-authors: Agnieszka Wylomanska
A bidimensional autoregressive model of order 1 with alpha-stable distributed noise is considered. Since in this case the classical measure of dependence known as the covariance function is not defined, the spatio-temporal dependence structure is described using the alternative measures, namely the codifference and the covariation functions. We investigate the asymptotic relation between these two dependence measures for the auto- and cross-dependence functions. We demonstrate the cases when the dependence measures are asymptotically proportional with the coefficient of proportionality equal to the stability parameter alpha. The theoretical results are supported by illustrating the asymptotic behavior of the dependence measures for an exemplary bidimensional alpha-stable AR(1) system.

E1694: **Bayesian sensitivity analysis of posterior predictive distribution in extreme value theory**
Presenter: **Jose Pablo Arias-Nicolas**, University of Extremadura, Spain
Co-authors: Eva Lopez Sanjuan, M Isabel Parra Arevalo
The objective of extreme value analysis is to model and measure tail events that occur with small probability, using only extreme values above some high threshold rather than using all the data. It is well known that, when we consider the values of the sample space above a certain value (threshold) the limit distribution function is a Generalized Pareto Distribution. Two measures that we find most useful and reliable for describing the tail of the distribution are Value at risk (VaR) and expected shortfall (ES). Under a Bayesian framework, we use the band distorted class to compute the range of these two risk measures for the posterior predictive density. The most important property is that the likelihood ratio order we are using in this class is preserved after the application the posterior belief. In the same way, we can see something similar for predictive distributions. We show that computations of sensitivity measures should be as easy as possible, possibly looking for the extremal distributions generating the class. Due to the fact that the distorted band depends on the election of the reference prior distribution (also on the distorted functions considered), we illustrate with some numerical examples, how this choice affects the calculation of the measures considered. Moreover, we compare the results considering the data from some standard populations: Normal, Exponential, Cauchy,...

E1776: **Goodness of fit test and model assessment for sparse multinomials**
Presenter: **Ioulia Papageorgiou**, Athens University of Economics and Business, Greece
The focus is on goodness-of-t tests under a very general framework where testing the fit of a model is equivalent with testing a hypothesis about the parameters from a multinomial distribution. A very realistic problem in such applications is that the sample size is small and at the same time the dimension of the problem is high, resulting to a sparse contingency table where asymptotic tests does not hold. Methods in the literature, such as resampling methods, posterior predictive p-values etc., require refitting of the model a number of times, and are time-consuming and computational effort demanding. Moreover, the distribution of the p-value is not uniform and therefore results are not interpretable. The proposed method is a variant of the posterior predictive p-value. The main advantage of the method is that no refitting of the model is required and the distribution of the resulting p-value is uniform. Consequently, the method outperforms all other competitive approaches with respect to computational efficiency and interpretability of the results. The proposed method is quite general and it can be implemented to a wide range of scientific areas. We present an application in capture-recapture models. More specifically, various models aiming to estimate animal population abundance based on a capture-recapture sample are assessed with respect to the model fitting, via the proposed methodology, and compared with previous relative results in the literature.

E1795: **Grouped portfolio optimization with pessimistic risk measure**
Presenter: **Sung Chul Hong**, University of Seoul, Korea, South
Co-authors: Eun Young Ko, Hosik Choi, Jong-june Jeon
In portfolio optimization theory recent studies of risk measures lead to a new strategy of the optimal asset allocation beyond the Markowitz mean-variance model. In particular, the α-risk receives much attention as one of the pessimistic risk measures. It is known that minimizing α-risk on portfolio management is closely related to the quantile regression model and thus various strategies of asset allocation can be derived from variants of quantile regression model. On this approach we propose an optimization method of portfolio that provides sparse and group-wise selection of assets based on α-risk. We formulate the problem as the regularized quantile regression model and solve our problem by ADMM algorithm. We

show the sparsity of asset allocation by numerical simulations and we also investigate out-of-sample performance in terms of various risk measures in real data analysis.

E1860: **Some characteristics of measurement invariance models and invariance indices**
Presenter: **Tomoya Okubo**, The National Center for University Entarance Examinations, Japan
The focus is on characteristics of some statistical models for measurement invariance analysis based on simulation studies. The measurement invariance sometimes occurs in international surveys, in which scores calibrated by the items do not ensure comparability among groups. Alignment method is a model that provide more practical information than conventional methods based on multi-group analysis. However, invariance index which is provided in the alignment method is not examined its statistical trait well. A purpose is to provide results of the simulation studies as well as real data analysis of international educational assessment data in order to reveal characteristics of some statistical models for measurement invariance analysis developed so far. Further, we propose new index for the measurement invariance analysis.

E1894: **A new characterisation-based test for symmetry**
Presenter: **Carl van Heerden**, North-West University, South Africa
Co-authors: Charl Pretorius
A new test for symmetry is proposed based on a lesser-known characterisation of symmetric distributions. We derive the limiting null distribution of the test and show that the test is consistent against general alternatives. The performance of the new test is evaluated and compared to that of existing tests by means of a Monte Carlo study. It is found that the newly proposed test performs favourably compared to the other tests.

E2029: **Coordinate representation of three-factorial compositional data**
Presenter: **Kamila Facevicova**, Palacky University Olomouc, Czech Republic
Co-authors: Karel Hron, Peter Filzmoser
Compositional data are commonly defined as positive vectors carrying relative information. Their relative nature prevents from applying standard statistical methods directly to raw compositions; instead, it is preferred to express compositional data in real logratio coordinates prior to their further processing. When a composition is formed according to more than just one factor, which results in so called compositional tables or cubes, the usual coordinate systems, designed primarily for vector compositional data, do not reflect the data structure sufficiently. The aim is to present an alternative coordinate representation of three-factorial compositional data - compositional cubes, which besides its favorable interpretation allows also to decompose the original data structure onto different sources of interactions between factors and to analyze these sources separately. The proposed methodology will be applied to real-world data and the possible use of spatial clustering and robust statistical methods will be discussed.

E2031: **kmlShape method to cluster longitudinal data**
Presenter: **Soonsun Kwon**, Ajou University, Korea, South
Longitudinal data arise where a response is observed on each subject repeatedly over the time. One possibility for the analysis of longitudinal data is to cluster them. The majority of clustering methods group together subject that have close trajectories at given time points. The methods group trajectories that are locally close, but not necessarily those that have similar shapes. In contrast, the kmlshape method is a longitudinal data partitioning algorithm based on the shapes of the trajectories, rather than on classical distances. We suggest the kmlShape method for high-dimensional datasets, and apply it to thyroidectomy cancer patients.

E2032: **bmeta: An R package for evidence synthesis using Bayesian meta-analysis**
Presenter: **Tao Ding**, University College London, United Kingdom
It is often the case that important research questions are studied more than once by different research teams at different locations and the outcomes from small studies can be diverse and conflicting. This may result in difficulties in decision making. However, combining available information from multiple sources to generate an integrated result may provide more indications. Meta-analysis is a commonly used statistical approach to realise this goal by integrating results from independent studies and is considered to play an essential role in evidence-based medicine. Since most applied implementations of meta-analysis are conducted under the Frequentist paradigm, there is the need to develop a package using Bayesian meta-analytic methods given its advantages (e.g. the observed data can be complemented by some prior belief and it takes fuller account of the uncertainties related to both parameter values and models). Therefore, the bmeta package for R is created. The b in the name stands for Bayesian.

E2039: **Bayesian uncertainty decomposition for hydrological projections**
Presenter: **Seonghyeon Kim**, Seoul National University, Korea, South
Co-authors: Ilsang Ohn, Yongdai Kim, Seung Beom Seo, Young-Oh Kim
There is a considerable uncertainty in a hydrological projection, which arisen from the multiple stages composing the hydrological projection. Uncertainty decomposition analysis evaluates contribution of each stage to the total uncertainty in the hydrological projection. Some uncertainty decomposition methods have been proposed, but they still have some limitations: (1) they do not consider nonstationarity in data and (2) they only use summary statistics of the projected data instead of the full time-series and lack a principled way to choose the summary statistic. We propose a novel Bayesian uncertainty decomposition method which can alleviate such problems. In addition, the proposed method provides probabilistic statements about the uncertainties. We apply the proposed method to the streamflow projection data for Yongdam Dam basin located at Geum River in South Korea.

CI020 Room Chancellor's Hall INVITED SESSION 2 Chair: Marcelo Fernandes

C0166: **Quantile forecast evaluation: An application to growth-at-risk**
Presenter: **Valentina Corradi**, University of Surrey, United Kingdom
A forecast evaluation procedure is introduced for multiple, possibly misspecified quantile models. The models are evaluated in terms of their relative unconditional coverage, that is we rank models in terms of the distance between actual and nominal coverage. The key novelty of our approach is that we do so uniformly over a compact set of quantile ranks, rather than at a single, pre-specified quantile level. In a final step, we then construct a model confidence set that contains all models which are 'equally good' over a weighted average of quantile ranks. As all model parameters are estimated using a recursive scheme, the contribution of recursive quantile estimation error has to be taken into account. Inference is based on block bootstrap p-values, and we establish a bootstrap Bahadur representation which is valid uniformly over quantile ranks and in a recursive setting. Finally, we apply our procedure to compare out-of-sample predictions for growth-at-risk of different quantile models and of professional forecasters.

C0167: **Price discovery and market microstructure noise**
Presenter: **Marcelo Fernandes**, Sao Paulo School of Economics, FGV, Brazil
Co-authors: Cristina Scherrer, Gustavo Dias
Using a continuous-time price discovery model, we show that the standard econometric framework typically yields inconsistent estimates of price discovery measures in the presence of market microstructure noise. We address this errors-in-variable issue using instrumental variables. We devise valid instruments for two alternative microstructure noise settings, and then establish the asymptotic behavior of the corresponding price discovery measures. We illustrate our findings by investigating price discovery for Alcoa, showing that market leadership conclusions depend heavily on whether we account or not for market microstructure noise.

C0168: **High dimensional influences and financial contagion**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Deniz Erdemlioglu, Stefano Nasini
The availability of high frequency information over cross-sectional dimensions, along with the disposal of important indicators (such as volatility, jumps and trading activity) constitutes a precious source of complex information, mirroring intrinsically interconnected aspects of the financial system. We rely on a novel network-based statistical approach to capture financial contagion from these integrated and complex data structures. The model relies upon influences across a large panel of equities, accounting for three distinct forms of interdependence: (1) the lagged interdependence, (2) the interdependence among different variables measured from high frequency data, and (3) the interdependence between industry-based equity clusters. The flexibility of our approach allows us to recover the pairwise influence structure, based on a specialized estimation approach for high-dimensional parameter spaces. Given the estimated influence matrix from the market data, we study interdependence dynamics among companies and perform additional economic analyses.

CO214 Room MAL B04 RISK MODELLING IN EQUITY AND OPTION MARKETS Chair: Jose Olmo

C0641: **Singular spectrum analysis for value at risk evaluation in stochastic volatilty models**
Presenter: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain
Co-authors: Javier Garcia
Estimation of the Value at risk (VaR) in Stochastic Volatility (SV) models requires prediction of the future volatility, which is a function of a latent variable that is not observable. In-sample and out-of-sample prediction of that unobservable variable is thus necessary. The former is related with signal extraction whereas the latter implies prediction of future values. Singular Spectrum Analysis (SSA) is a useful tool for both purposes. The focus is mainly on out-of-sample predictions, comparing the performance of SSA with other forecasting techniques when used for VaR evaluation in SV models. Their empirical performance is also analysed in a daily series of SP500 returns.

C0985: **Optimal quantitative risk factors**
Presenter: **Richard McGee**, University College Dublin, Ireland
Co-authors: Richard McGee, Jose Olmo
Traditional quantitative risk factors are constructed using equally-weighted or value-weighted long-short portfolios from ranked, sorted quantiles of the cross section of stocks. The weighting scheme within the quantile portfolios is therefore agnostic of the target relationship between the candidate attribute and stock returns. We propose a new approach to constructing factor mimicking portfolios. The optimal weightings for each stock within the portfolio are informed by the relationship between the candidate attribute and expected stock returns. We evaluate the new factors empirically for the size and book to market stock characteristics.

C1096: **On market price extremes of underlying shares in the options market**
Presenter: **Marie Kratz**, ESSEC Business School, CREAR, France
Different quantitative methods are examined to recover the risk neutral distribution function associated with the prices of option on bank shares. After having assessed their qualities by recovering market option prices from these distributions, we compare Value-at-Risk implied by the estimated risk neutral distribution to historical Value-at-Risk of the share prices. Option prices of American banks designated as SIFIs (Systemic Financial Institutions) are considered. We show that, contrary to what is to be expected for a market composed of risk averse investors, the implicit VaR is under-estimated compared to the one obtained from real data in normal times, while over-estimated in times of crisis. Consequences of this result will be discussed to monitor market sentiments.

C1846: **Granger-causality detection in high-dimensional systems using feedforward neural networks**
Presenter: **Jose Olmo**, University of Southampton, United Kingdom
Co-authors: Hector Calvo-Pardo, Tullio Mancini
A novel methodology is proposed to detect Granger causality in mean using feedforward neural networks. The approach accommodates unknown dependence structures between the elements of highly-dimensional multivariate time series with weak and strong persistence characterized by a vector of lags potentially increasing to infinity. To do this, we propose a two-stage procedure. In a first stage, we fit a neural network given by an optimal number of nodes in the intermediate hidden layers. This is done by minimizing the entropy of the neural network - maximizing the transfer of information between input and output variables. In a second stage, we apply a novel sparse double group lasso penalty function to identify the variables that have predictive ability in the multivariate time series and, hence, Granger cause of the others. The penalty function inducing sparsity is applied to the weights characterising the nodes of the neural network and allows us to add interpretability to the neural network by mapping the nodes of the neural network to the variables exhibiting Granger causality. We show the correct identification of these weights when the number of variables and lags is finite and also when increases to infinity with the sample size. An application to the recently created Tobalaba network of renewable energy companies shows how to exploit Granger-causality for uncovering the role of the net in increasing connectivity between companies and forecast ability.

CO396 Room MAL B20 ECONOMICS OF CRYPTOCURRENCIES Chair: Marco Lorusso

C0741: **Will Bitcoin mining lead to global environmental catastrophe?**
Presenter: **Ladislav Kristoufek**, Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic
Bitcoin as a major cryptocurrency has come up as a shooting star of the 2017 and 2018 headlines. After exploding its price twenty times just in the twelve months of 2017, the tone has changed dramatically in 2018 after major price corrections and increasing concerns about its mining power consumption and overall sustainability. The dynamics and interaction between Bitcoin price and its mining costs have become of major interest. We show that these two quantities are tightly interconnected and they tend to a common long-term equilibrium. Mining costs adjust to the cryptocurrency price with the adjustment time of several months up to a year. Current developments suggest that we have arrived at a new era of Bitcoin mining where marginal (electricity) costs and mining efficiency play the prime role. Forecasting models building on this dynamics suggest that even though the hashrate will keep increasing (as well as Bitcoin price), it will be controlled by increasing mining efficiency and reward halving. In effect, the maximum profitable electricity price will keep decreasing forcing the use of cheap/renewable sources of energy. Environmental catastrophe induced by Bitcoin mining thus does not seem likely to threaten us.

C1341: **Another look at cryptocurrency bubbles**
Presenter: **Marc Gronwald**, University of Aberdeen, United Kingdom
Cryptocurrency bubbles are considered. First, it points out that a number of recent papers on cryptocurrency bubbles are flawed due to an insufficient consideration of the fundamental value of cryptocurrencies. As even fiat money is said to exhibit features of bubbles, the same applies to cryptocurrencies. Thus, any empirical investigation into either the presence of cryptocurrency bubbles or the fundamental value of cryptocurrencies is needless. Second, a short empirical analysis into the relationship of the prices of Etherum and Bitcoin is conducted. Evidence of explosive periods is found in the price of Etherum even if this price is expressed in terms of Bitcoin rather than US Dollars. These periods, however, are found to be in the first half of 2016 and 2017, respectively, but not during the price peak period of Bitcoin witnessed end of 2017 and beginning of 2018.

C0503: **Comovement and bubbles in cryptocurrency markets**
Presenter: **Pierangelo De Pace**, Pomona College, United States
Co-authors: Jayant Rao

Using daily data spanning the period between the end of April 2013 and the end of November 2018, we analyze the correlations of daily price returns for nine major cryptocurrencies and estimate their time evolutions by means of both a bivariate and multivariate modelling approach. We detect pronounced time variation in the comovement of price returns and find it to be generally increasing for all pairs of cryptocurrencies between early 2017 and late 2018. Furthermore, we adopt a recently developed right-tail variation of the Augmented Dickey-Fuller unit root test to identify and date-stamp periods of mildly explosive behavior (bubbles) in the time series of the Network Value to Transactions (NVT) ratio of six of these cryptocurrencies and show statistically significant evidence of mildly explosive dynamics in all of them. Bubbles are not necessarily synchronized across cryptocurrencies. However, in 2018, most major cryptocurrencies experience significant simultaneous distress associated with quickly rising NVT ratios. Instability and distress appear to be a steady feature of cryptocurrency markets. Our results suggest that Bitcoin may be leading the way and dragging other major cryptocurrencies.

C0697: **A new economic framework: A DSGE model with cryptocurrency**
Presenter: **Marco Lorusso**, Northumbria University, United Kingdom
Co-authors: Francesco Ravazzolo, Stylianos Asimakopoulos

A Dynamic Stochastic General Equilibrium (DSGE) model is developed to evaluate the economic repercussions of cryptocurrencies. We assume that the representative household maximizes its utility given by consumption, leisure and both government currency and cryptocurrency holdings. Our model includes entrepreneurs that determine the supply of cryptocurrency in the economy. We also consider a central bank setting the nominal interest rate following a general augmented Taylor-type interest-rate rule. In particular, the nominal rate responds not only to the interest rate in the previous period and to deviations of output and inflation from their steady-state values, but also to nominal money growth in government currency and cryptocurrency. We estimate our model with Bayesian techniques using US monthly data for the sample period 2013:M6-2019:M3. Our impulse response analysis shows the effects of a traditional shock to household's demand for real balances of government currency as well as to a new shock to household's demand for real balances of cryptocurrency. Moreover, we evaluate the response of main macroeconomic fundamentals to productivity shocks for production of cryptocurrency.

CO649 Room MAL B35 INFLATION DYNAMICS Chair: Edward Knotek

C0572: **Real-time density nowcasts of U.S. inflation: A model-combination approach**
Presenter: **Edward Knotek**, Federal Reserve Bank of Cleveland, United States
Co-authors: Saeed Zaman

A model combination framework is developed to produce density nowcasts for U.S. headline and core inflation at a trading-day frequency. We apply a flexible aggregation strategy and several weighting schemes to combine individual density nowcasts coming from three classes of mixed-frequency inflation nowcasting models. We assess the properties of the density nowcasts generated from the individual classes of models and from the grand combination of the three model classes using high-frequency, real-time data over the period 2000-2015. As information accumulates over a month or quarter, density nowcast accuracy steadily improves as judged by predictive scores, but only the grand combination is always among the most accurate and comes closest to passing all the necessary tests to be correctly calibrated. Both point and density nowcasts from the grand combination outperform survey benchmarks in the case of headline inflation.

C0693: **Does money growth predict inflation: Evidence from vector autoregressions using four centuries of data**
Presenter: **Par Osterholm**, Orebro university, Sweden
Co-authors: Rodney Edvinsson, Sune Karlsson

New evidence is added to a long-debated macroeconomic question, namely, whether monetary aggregates have predictive power for inflation or, put differently, whether monetary aggregates Granger cause inflation. We study this issue by employing vector autoregressive models to unique data. Using a historical dataset - consisting of annual Swedish data on money growth and inflation ranging from 1620 to 2018 - we conduct analysis both within- and out-of-sample. Using state-of-the-art Bayesian methods, we estimate models with stochastic volatility and also allow for the possibility of drifting parameters. Model selection based on marginal likelihoods indicates that a model with both stochastic volatility and drifting parameters is preferred. Concerning the issue of Granger causality, within-sample analysis provides strong evidence in favour of money growth Granger causing inflation when estimated on the full sample. Our out-of-sample forecast exercise does, however, not support this picture; the point forecast accuracy from the model where inflation is exogenous with respect to money growth tends to be on par with that from the model where it is endogenous. There is, however, a tendency for the model in which inflation is endogenous with respect to money growth to generate better density forecasts.

C0777: **Forecasting US inflation in real-time**
Presenter: **Chad Fulton**, Federal Reserve Board of Governors, United States
Co-authors: Kirstin Hubrich

A real-time forecasting exercise for US inflation is performed, investigating whether and how additional information - additional macroeconomic variables, expert judgement, or forecast combination - can improve forecast accuracy and robustness. This is particularly relevant now, as the length of the current expansion nears that of the longest on record and unemployment has fallen to a rate not seen for nearly sixty years. Distinguishing features of our study include a focus on forecasting performance before, during, and after the global financial crisis and the use of (published) Federal Reserve Board staff forecasts contained in Tealbooks. We find that (1) while simple models remain hard to beat, model-based forecasts of the types we consider can improve on them, especially in the post-crisis period; (2) judgmental forecasts are competitive and can outperform even simple benchmark models; (3) aggregating forecasts of inflation components can improve performance compared to forecasting the aggregate directly; (4) forecast combination approaches provide competitive forecasts and robustify against bad forecasts.

C1326: **Forecasting with unknown unknowns: Censoring and fat tails on the Bank of England's monetary policy committee**
Presenter: **James Mitchell**, University of Warwick, United Kingdom
Co-authors: Martin Weale

The production and evaluation of density forecasts is considered by paying attention to if and how the probabilities of outlying observations are quantified and communicated. Particular focus is given to the 'censored' nature of the Bank of England's fan charts, given that - which is commonly ignored - they describe only the inner 90% (best critical region) of the forecast distribution. A new estimator is proposed that fits a potentially skewed and fat tailed density to the inner observations, acknowledging that the outlying observations may be drawn from a different but unknown distribution. In forecasting applications, motivation for this could reflect the view that outlying forecast errors reflect (realised) unknown unknowns or events not expected to recur that should be censored before quantifying known unknowns.

CO246 Room MAL B36 UNCERTAINTY AND TEXT ANALYSIS Chair: Svetlana Makarova

C0811: **Economic policy uncertainty and stock market participation**
Presenter: **Eniko Gabor-Toth**, Deutsche Bundesbank, Germany

Does economic policy uncertainty affect household stockholding? To answer this question, a novel measure is created for household exposure to economic policy uncertainty news by combining survey information on the hours a household spends in reading newspapers and the frequency of such news in the popular press during a households pre-interview period. After controlling for household fixed effects, month-year fixed effects and time-varying cognitive skills, we find that households with more exposure to economic policy uncertainty news are less likely to invest in stocks directly or through mutual funds. This effect is independent of the VIX and household stock-price expectations.

C1558: **Newspaper-based economic uncertainty indices for Poland**
Presenter: **Marcin Holda**, National Bank of Poland, Poland

Using text mining and web scraping techniques, we develop newspaper-based economic uncertainty measures for Poland. We build general economic and economic-policy uncertainty indices, as well as category-specific ones designed to capture e.g. the economic uncertainty related to fiscal policy or to stockmarket movements. Several types of evidence suggest that these indices do proxy for changes in economic uncertainty in Poland. In particular, our measures spike around uncertainty-laden events or periods, such as the initial phase of Poland's post-communist economic transition, the global financial crisis or the European debt crisis that followed. Our indices also exhibit correlation with a variety of other indicators of economic uncertainty, such as financial-market data and results of corporate surveys. The newspaper-based indices behave similarly to uncertainty indicators developed using other textual data and are strongly correlated with relevant economic uncertainty indicators developed by other researchers.

C0912: **Components of uncertainty**
Presenter: **Vegard Larsen**, Norges Bank, Norway

Uncertainty is acknowledged to be a source of economic fluctuations. But, does the type of uncertainty matter for the economy's response to an uncertainty shock? A novel identification strategy is offered to disentangle different types of uncertainty. It uses machine learning techniques to classify different types of news instead of specifying a set of keywords. It is found that, depending on its source, the effects of uncertainty on macroeconomic variable may differ. Both good (expansionary effect) and bad (contractionary effect) types of uncertainty exist are found.

C0915: **Country-specific uncertainty indices and machine learning: The case of Russia**
Presenter: **Svetlana Makarova**, University College London, United Kingdom
Co-authors: Wojciech Charemza, Krzysztof Rybinski

Problems related to the construction of country-specific geopolitical and economic policy uncertainty indices have been identified and analysed. The methodology is based on the textual analysis of data extracted from local language newspapers through unsupervised machine learning. The problems include (1) identification of the external and internal (idiosyncratic) factors affecting uncertainty; (2) crowding out and covering up the uncertainty-related topics by other news items; (3) identification of country-specific effects of global geopolitical and economic policy uncertainties. Problems (1) and (2) have been tackled by using the Latent Dirichlet Allocation algorithms with various settings to recognise economic and policy-related topics and applying Word2Vec model to categorise uncertainty-related terms. For problem (3), quantile correlation techniques have been applied to find a relationship between changes in the global (worldwide) uncertainty and country-specific local language indices. The data used are collected from four main Russian newspapers and cover the period from 1992 to 2018.

CO402 Room Gordon SEMI- AND NONPARAMETRIC REGRESSION FOR TIME SERIES AND PANEL DATA I Chair: Joachim Schnurbus

C1632: **Nonparametric multi-dimensional fixed effects panel data models**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Daniel Henderson, Juan Manuel Rodriguez-Poo

Multi-dimensional panel data sets are routinely employed to identify marginal effects in empirical research. Fixed effects estimators are typically used in order to deal with potential correlation between unobserved effects and regressors. Nonparametric estimators for one-way fixed effects models exist, but are cumbersome to employ in practice as they typically require iteration, marginal integration or profile estimation. We develop nonparametric estimators for the gradient of the conditional mean that work for essentially any dimension fixed effects model, have closed-form solutions and can be estimated in a single-step. Cross-validation bandwidth selection procedures are proposed and the asymptotic properties (for a fixed or large time dimension) of our estimators are given. Finite sample properties are shown via simulations, as well as with an empirical application which further extends our estimators to the partially linear setting.

C1625: **Inference for multidimensional nonparametric panel data models**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Alexandra Soberon

Inference for multidimensional nonparametric panel data models is addressed. We consider tests for correct parametric specification, tests for relevance of variables, and Hausman-type tests for fixed versus random effects. We are able to show the asymptotic distribution of said tests, propose and show validity of bootstrap procedures and show the finite sample properties.

C1476: **Semi- and nonparametric modeling of environmental time series distributions**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Harry Haupt

Nonparametric regression-based approaches are proposed for modeling the distribution of stochastic processes which may exhibit nonlinearities and non-stationarities driven by trend and/or seasonal patterns. Special emphasis lies on a novel approach for smoothing of potentially multiple seasonal structures. An application to environmental time series reveals that the smoothing approach provides a simple and computationally cheap approach for distribution estimation and forecasting of such processes.

C1485: **Practical aspects of using nonlinear moment conditions in linear dynamic panel data models**
Presenter: **Markus Fritsch**, University of Passau, Germany
Co-authors: Joachim Schnurbus, Andrew Adrian Yu Pua

The focus is on the estimation of the lag parameter of linear dynamic panel data models with first order dynamics based on nonlinear (quadratic) moment conditions. The contribution is twofold: First, we show that extending the standard assumptions by mean stationarity and time series homoscedasticity and employing these assumptions in estimation restores standard asymptotics and mitigates the non-standard distributions found in the literature. Second, we consider an IV estimator based on the quadratic moment conditions that consistently identifies the true population parameter under standard assumptions. Standard asymptotics hold for the estimator when the cross section dimension is large and the time series dimension is finite. We also suggest a data-driven approach to obtain standard errors and confidence intervals that preserves the time series dependence structure in the data.

CO864 Room Montague TOPICS IN FINANCIAL ECONOMETRICS II Chair: Leopold Soegner

C0783: **Efficient Bayesian estimation of the stochastic volatility model with leverage**
Presenter: **Darjus Hosszejni**, WU Vienna University of Economics and Business, Austria
Co-authors: Gregor Kastner
The sampling efficiency of MCMC methods in Bayesian inference for stochastic volatility (SV) models is known to highly depend on the actual parameter values, and the effectiveness of samplers based on different parameterizations differs significantly. We derive novel samplers for the centered and the non-centered parameterizations of the practically highly relevant SV model with leverage, where the return process and the innovations of the volatility process are allowed to correlate. Moreover, based on the idea of ancillarity-sufficiency interweaving, we combine the resulting samplers in order to achieve superior sampling efficiency, irrespective of the baseline parameterization. The method is implemented using R and C++.

C1002: **Forecasting benchmarks of long-term stock returns via machine learning**
Presenter: **Michael Scholz**, University of Graz, Austria
Co-authors: Jens Perch Nielsen, Ioannis Kyriakou, Parastoo Mousavi
Recent advances in pension product development seem to favour alternatives to the risk free asset often used in the financial theory as a performance standard for measuring the value generated by an investment or a reference point for determining the value of a financial instrument. To this end, we apply the simplest machine learning technique, namely, a fully nonparametric smoother with the covariates and the smoothing parameter chosen by cross-validation to forecast stock returns in excess of different benchmarks, including the short-term interest rate, long-term interest rate, earnings-by-price ratio, and the inflation. We find that, net-of-inflation, the combined earnings-by-price and long-short rate spread form our best-performing two-dimensional set of predictors for future annual stock returns. This is a crucial conclusion for actuarial applications that aim to provide real-income forecasts for pensioners.

C1439: **A subspace estimator for mixed frequency VAR systems**
Presenter: **Philipp Gersing**, Vienna University of Technology, Austria
Co-authors: Manfred Deistler
A novel estimation procedure is proposed for mixed frequency data where the underlying high frequency process is a VAR process. We derive an explicit minimal stable and miniphase state space representation for a blocked process which second moments correspond to all those second moments observable under mixed frequency. Based on this representation a relatively simple subspace estimator for the high frequency parameters is derived which makes full use of the information available and is consistent and efficient. The computational burden of the procedure is much lower compared to methods involving Kalman filtering. We aim to extend the estimator to the case of mixed frequency generalized dynamic factor models where the unobserved component has rational spectral density.

C0282: **Forecasting the realized variance in the presence of intraday periodicity**
Presenter: **Rodrigo Hizmeri**, Lancaster University, United Kingdom
Co-authors: Ana-Maria Dumitru, Marwan Izzeldin
The impact of intraday periodicity on forecasting realized volatility is examined by using a heterogeneous autoregressive model (HAR) framework. We show that periodicity inflates the variance of the realized volatility and biases jump estimators. This combined effect adversely affects forecasting. To account for this, we propose a periodicity-adjusted model, HARP, where predictors are built from the periodicity-filtered data. We demonstrate empirically (using 30 stocks from various business sectors and the SPY for the period 2000–2016) and via Monte Carlo simulations that the HARP models produce significantly better forecasts, especially at the 1-day and 5-days ahead horizons.

CO232 Room Woburn MACROECONOMIC POLICY Chair: Michael Owyang

C0609: **On the role of trade and production networks for the effectiveness of government spending**
Presenter: **Nora Traum**, HEC Montreal, Canada
The purpose is to study the effects of government spending shocks in an open economy with multiple interconnected production sectors that differ in terms of factor intensities, imported and domestic intermediate inputs, the degree of price rigidities, and contribution to GDP. The model is fit to U.S. data. While previous studies have emphasized the importance of heterogeneity in accounting for the effectiveness of sector-specific public spending shocks in a closed economy, we focus on the importance of the source of government spending and its interaction with value added chains for understanding the quantitative effectiveness of targeted discretionary spending.

C1055: **Revisiting the Fed's forecasting advantage**
Presenter: **Amy Guisinger**, Lafayette College, United States
Co-authors: Michael Owyang, Michael McCracken
Previous studies have found that the Federal Reserve's (Greenbook) forecasts of inflation are superior to that of professional forecasters. More recently, however, it has been suggested that this advantage has been dissipating over time. We revisit previous findings exploiting the additional observations and confirm the Federal Reserve forecasting advantage for inflation, GDP, and the unemployment rate. We then investigate the origin of the Fed's forecasting advantage, taking into account that professional forecasters appear to be catching up with the Fed. We focus on an explanation that was dismissed previously, that the Fed's forecast are conditional on the path of policy, about which the Fed may have more information. To do this, we examine whether the Federal Reserve's advantage remains after controlling for information about future monetary policy. We then consider whether the Fed's advantage remains when the market is operating under the same (or similar) beliefs as the Fed.

C1483: **Unconventional monetary policy, (a)synchronicity and the yield curve**
Presenter: **Karlye Dilts Stedman**, Federal Reserve Bank of Kansas City, United States
International spillovers are examined from unconventional monetary policy between the US, Euro area, UK and Japan, focusing on the effect of exit timing on the term structure of interest rates. Using high-frequency futures data to identify monetary policy surprises and controlling for contemporaneous news, we find that spillovers increase during the period of asynchronous policy normalization, wherein the term premium gains importance in driving spillovers compared to previous periods. Local projections suggest persistent spillovers from the Federal Reserve, whereas other spillovers fade quickly. Through the lens of a shadow rate term structure model (SRTSM), we find that these surprises elicit, domestically and internationally, revisions to both the expected path of short-term interest rates and required risk compensation.

C1998: **Impact of foreign official purchases of U.S. treasuries on the yield curve**
Presenter: **Erin Wolcott**, Middlebury College, United States
Foreign governments went from owning ten percent of publicly-held U.S. Treasury debt in 1985 to owning the majority in 2008. Since the financial crisis, foreign governments have reduced their Treasury positions. The foreign official purchases are revealed to have depressed short- and medium-term interest rates in the U.S., despite conventional wisdom pointing towards the long end of the yield curve. We estimate a Gaussian affine term structure model, augmented with a vector autoregression of macroeconomic variables, to examine effects over the entire yield curve, as opposed to a single maturity. Distinguishing which part of the yield curve foreign official purchases move is important for monetary policy. If segments of the yield curve are increasingly determined by international financial markets, then it may be more difficult for the Federal Reserve to implement its

interest rate policy and mitigate short-term fluctuations.

CO254 Room MAL 153 MULTIVARIATE QUANTILE MODELS Chair: Gabriel Montes-Rojas

C0427: **On the distribution of impulse-response functions in macroeconomic shocks**
Presenter: **Gabriel Montes-Rojas**, Universidad de Buenos Aires, Argentina

A multivariate vector autoregressive model is used to construct the distribution of the impulse-response functions of macroeconomics shocks. In particular, the aim is to study the distribution of the short, medium and long term effects of shocks and evaluate the occurrence of extreme events. The model considers a reduced form quantile vector autoregressive model where heterogeneity in conditional effects can be evaluated using a simulation of uniformly distributed random vectors each period. The proposed model provides point estimation of the entire distribution of potential events. An empirical example on evaluating monetary shocks is presented.

C0686: **Nonparametric estimation of the variance function in an explosive autoregressive model**
Presenter: **Yang Zu**, University of Nottingham, United Kingdom
Co-authors: Dave Harvey, Steve Leybourne

Nonparametric estimation is considered for the innovation variance function in an autoregressive model that can exhibit unit root, explosive and stationary regimes, allowing for behaviour often seen in financial data where bubble and crash episodes are present. The model permits multiple regime changes occurring at unknown points in time. Extant variance function estimators lack consistency for our model. We thus propose a new truncation-based kernel smoothing estimator, which we show is uniformly consistent for the innovation variance function. In a Monte Carlo simulation, we study the finite sample performance of our estimator and highlight the role of truncation in increasing the variance estimation accuracy.

C0849: **On the unbiased asymptotic normality of quantile regression with fixed effect**
Presenter: **Antonio Galvao**, University of Arizona, United States

Nonlinear panel data models with fixed individual effects provide an important set of tools for describing microeconometric data. In a large class of such models (including probit, proportional hazard and quantile regression to name just a few) it is impossible to difference out the individual effects, and inference is usually justified in a 'large n large T' asymptotic framework. However, there is a considerable gap in the type of assumptions that are currently imposed in models with smooth score functions (such as probit, and proportional hazard) and quantile regression. We show that this gap can be bridged and establish asymptotic unbiased normality for fixed effects quantile regression panels under conditions on n and T that are very close to what is typically assumed in standard nonlinear panels. The results considerably improve upon existing theory and show that quantile regression is applicable to the same type of panel data (in terms of n, T) as other commonly used nonlinear panel data models. Thorough numerical experiments confirm our theoretical findings.

C0189: **Partially linear quantile regression model with time-varying loadings**
Presenter: **Alev Atak**, City University London, United Kingdom
Co-authors: Gabriel Montes-Rojas, Yonghui Zhang, Jose Olmo

A semiparametric quantile regression model with factor-augmented predictors and time-varying factor loadings is developed. We propose a two-stage procedure. In the first step, we estimate factors from the mean regression model using a local version of the principal component method and we construct an average quantile regression. In the second step, we obtain partially linear varying coefficient quantile regression using the estimated factors derived in the first step. The proposed method extracts and combines distributional information across different probability masses. Uniform consistency and weak convergence of the estimated quantile factor loading processes are established under general assumptions. We evaluate the volatility-return relationship in real-time applications by observing the behavior of time-varying factor loadings in lower, mid and upper quantiles. We find strong evidence of heterogeneity in dynamic responses.

CC822 Room MAL B02 CONTRIBUTIONS IN ASSET PRICING Chair: Thomas Conlon

C0394: **Discriminating between GARCH models for option pricing by their ability to compute accurate VIX measures**
Presenter: **Christophe Chorro**, University, France

The pricing performances of a large collection of GARCH models is discussed by questioning the global synergy between the choice of the affine/non-affine GARCH specification, the use of competing alternatives to the Gaussian distribution, the selection of an appropriate pricing kernel and the choice of different estimation strategies based on several sets of financial information. Furthermore, an important question in relation to the correlation between the performance of a pricing scheme and its ability to forecast VIX dynamics is answered. VIX analysis clearly appears as a parsimonious first-stage filter to discard the worst GARCH option pricing models.

C1693: **Horizon-specific risks, higher moments, and asset prices**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic
Co-authors: Jozef Barunik

Asset pricing traditionally works with information aggregated over horizons, however investors preferences are horizon-specific. Decomposing returns, and risk factors to components representing individual horizons may hence provide valuable insights into pricing mechanisms of investors. With increasing size of factor-investing literature, the number of factors approximating risk, and possibly explaining the cross section of returns is growing rapidly. However, most of the factors perform poorly in subsequent out-of-sample testing. Therefore, attention should be turned to theory-based factors approximating the risks such as moments of the return distribution that are found to be priced empirically. We derive an asset pricing model that contains second, third and fourth centralized moments of returns on aggregate wealth decomposed to short-run, medium-run, and long-run components. The horizon-specific model outperforms CAPM, and Four-Moment CAPM in explaining the mean returns of S&P 500 stocks, and Exchange Traded Funds. Moreover, the derivation assuming a general utility function of wealth allows us to quantify the discrepancies in investors' risk tastes like risk aversion or prudence in different horizons.

C1808: **Dynamic quantile model for bond pricing**
Presenter: **Frantisek Cech**, Charles University, Czech Republic
Co-authors: Jozef Barunik

A dynamic quantile model for bond pricing is introduced. We consider an agent who values securities by maximizing quantile level of her utility instead of the expected utility function. The shift to quantile preferences from the traditional von Neuman-Morgensterns utility allows us to study the bond pricing by the economic agents differing at their risk aversion. In the empirical application, we focus on the various US and German government bonds. We rely on the flexible quantile regression framework applied to the bond futures contract from the CBOT and EUREX exchanges. We find idiosyncratic risk to price government bond futures at daily frequency, although, the pricing patterns differ between US and Germany. At the lower frequency, we identify instantaneous forward rates to serve as valid risk factors at several quantiles for US government bond futures.

C0334: **Global bond market interaction: An arbitrage-free dynamic Nelson Siegel modeling approach**
Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

A previous approach is extended to an arbitrage-free setting, proposing a global factor model in which country yield-level and slope factors may

depend on global-level, slope and curvature factors as well as country-specific factors. Using a monthly dataset of government bond yields for Germany, Japan, the US, and the UK from January 1995 to November 2018, we extract global and country-specific factors for both the full sample. The results indicate strongly that global yield-level, slope, and curvature factors do indeed exist and are economically important, accounting for a significant fraction of variation in country bond yields with interesting differences across countries. Moreover, the global yield factors appear linked to global macroeconomic fundamentals. We show, in particular, that curvature factors are key to explaining term premium dynamics, and appear more important in the second sub-sample.

CC821 Room MAL 351 CONTRIBUTIONS IN PORTFOLIO OPTIMIZATION I Chair: Abraham Lioui

C1605: **Dynamic Black Litterman copula based optimal portfolios with tail constraints**
Presenter: **Maziar Sahamkhadam**, Linnaeus University, Sweden
Co-authors: Andreas Stephan, Ralf Ostermark

The original Black-Litterman (BL) approach assumes normality, constant conditional distribution and no tail dependency, neither symmetric nor asymmetric. We estimate returns conditional distribution from a dynamic BL approach and model the tail dependency by applying truncated regular vine (Rvine) copula. Furthermore, reward-risk ratio optimizations generally consider only two portfolio characteristics, expected return and risk. It is shown that including tail constraints leads to more flexible portfolio strategies, combining tail and classical risk-return optimization techniques. Conditional Value-at-Risk (CVaR) is used as downside risk measure and added in classical reward-risk optimization. To examine the performance of the suggested forecasting models and optimization techniques, we perform out-of-sample back-testing for several portfolio strategies applied to a data set consisting of 30 stocks listed on the Stockholm exchange. We compare the results with benchmark portfolios including equally weighted (EQW) portfolio and portfolios obtained from dynamic BL model with out copulas. The results show more flexibility and frequent out-performance for the tail constraint augmented portfolios. In general, the suggested version of BL approach outperforms the benchmark models regarding both portfolio return and risk measures.

C1486: **Stochastic dominance inefficiency tests**
Presenter: **Sofia Anyfantaki**, Athens University of Economics and Business and Bank of Greece, Greece
Co-authors: Nikolas Topaloglou, Esfandiar Maasoumi, Jue Ren

The most common approach to test for stochastic dominance is to posit the null hypothesis of dominance. Rejection of the null, however, does not imply dominance since it can also happen that the test fails to rank the two distributions. A statistical test is proposed for the stochastic dominance efficiency of a given portfolio when the null hypothesis is inefficiency. An analytical characterization of stochastic dominance inefficiency and the null limit distribution for the associated empirical test statistic are derived. Feasible approaches to statistical inference based on bootstrapping and numerical optimization are developed. The test is used to empirically establish whether well known mutual funds are efficient during different market conditions with respect to (1) all possible portfolios constructed from a set of 248 funds and (2) different market proxies.

C1606: **A new information criterion for the Sharpe ratio**
Presenter: **Jakob Soehl**, TU Delft, Netherlands
Co-authors: Dirk Paulsen

When the in-sample Sharpe ratio is obtained by optimizing over a k-dimensional parameter space, it is a biased estimator for what can be expected on unseen data (out-of-sample). We derive (1) an unbiased estimator adjusting for both sources of bias: noise fit and estimation error. We then show (2) how to use the adjusted Sharpe ratio as model selection criterion analogously to the Akaike Information Criterion (AIC). Selecting a model with the highest adjusted Sharpe ratio selects the model with the highest estimated out-of-sample Sharpe ratio in the same way as selection by AIC does for the log-likelihood as measure of fit.

C1660: **Sampling distributions of optimal portfolio weights and characteristics in low and large dimensions**
Presenter: **Erik Thorsen**, Stockholm University, Sweden
Co-authors: Taras Bodnar, Nestor Parolya, Holger Dette

Optimal portfolio selection problems are determined by the (unknown) parameters of the data generating process. If an investor want to realise the position suggested by the optimal portfolios he/she needs to estimate the unknown parameters and account for the parameter uncertainty introduced into the decision process. Most often, the parameters of interest are the population mean vector and the population covariance matrix of the asset return distribution. We characterise the exact sampling distribution of the estimated optimal portfolio weights and their characteristics by deriving their sampling distribution which is present in terms of a stochastic representation. This approach possesses several advantages, like (i) it determines the sampling distribution of the estimated optimal portfolio weights by expressions which could be used to draw samples from this distribution efficiently; (ii) the application of the derived stochastic representation provides an easy way to obtain the asymptotic approximation of the sampling distribution. The later property is used to show that the high-dimensional asymptotic distribution of optimal portfolio weights is a multivariate normal and to determine its parameters. Via an extensive simulation study, we investigate the finite-sample performance of the derived asymptotic approximation and study its robustness to the violation of the model assumptions used in the derivation of the theoretical results.

CC828 Room MAL 352 CONTRIBUTIONS IN EMPIRICAL MACROECONOMICS Chair: Damjan Pfajfar

C1896: **Estimating the effects of the Eurosystem's asset purchase programme at the country level**
Presenter: **Michael Scharnagl**, Deutsche Bundesbank, Germany
Co-authors: Martin Mandler

The macroeconomic effects of the Eurosystems asset purchases on the four largest euro area economies are assessed by using simulation exercises which combine unconventional monetary policy shocks with a fixed policy rate for the duration of the purchase programme. We identify unconventional monetary policy shocks in a large Bayesian vector autoregressive (BVAR) model as shocks to the term structure of interest rates using zero and sign restrictions. We propose a multi-country model in which we impose identification assumptions mainly on euro area aggregate financial variables and on country averages of output and price responses. Furthermore, the multi-country structure allows testing for cross-country differences in the effects of the asset purchase programme in a statistically rigorous way using the posterior of the difference between the country-specific effects. We estimate positive output effects in all countries as well as positive effects on bank lending to firms. Effects on HICP inflation, generally, are much weaker. We find substantial cross-country heterogeneity with the largest price level effects in Spain while output effects were smallest in France and inflation effects were smallest in Italy.

C1347: **Investigating the determinants of exchange rate stability**
Presenter: **Essam Atta Arsanious Ghaprial**, Banque Misr, Egypt
Co-authors: Ahmed Mabrouk

The major drivers of nominal and real effective exchange rates foreign exchange rate are investigated in a sample of approximately 30 countries between 1977 and 2017. We used multiple linear regression analysis by Fixed GLS AR1, and Pooled Mean Group (PMG). The main drivers of nominal exchange rate in the long run are found to be: broad money, interest rate, claims on central government, government final consumption expenditures, trade, inflation, GDP per capita, oil prices, school enrollment and the expected exchange rate in the long run. While in the short run it is effected mainly by interest rate, general government final consumption expenditure, inflation, taxes, and the expected exchange rate. On the other hand, the main drivers of real effective exchange rate in the long run are found to be: current account balance, gross capital formation, trade,

school enrollment, oil prices, total reserves, broad money, interest rate, claims on central government, government final consumption expenditures, GDP per capita, taxes, and the expected exchange rate. While in the short run it is only trade, unemployment rate, and the expected exchange rate.

C0445: **Low inflation in advanced economies**
Presenter: **Luis J Alvarez**, Bank of Spain, Spain
Co-authors: Ana Gomez-Loscos, Lola Gadea

In recent years, actual inflation in most advanced economies has been below Central Bank's targets despite very expansionary monetary policies and a recovery in economic activity. This protracted period of low inflation has affected not only headline but also core inflation. We use a sample 27 advanced economies and a detailed sectoral breakdown to account for product heterogeneity in price setting. We find clear evidence of breaks in mean inflation and estimate two-state Markov switching models. We clearly distinguish between low and high inflation states and find that the conditional probability of remaining in low inflation is high. Furthermore, country inflation dynamics shows some common features, so we use Finite Mixtures Markov switching models to endogenously identify groups of countries. We find that non-European countries are clearly split from European ones and, in Europe, we distinguish core (those with price stability, such as Germany) and peripheral countries (traditionally inflationary countries, such as Spain). Finally, we estimate Markov switching models with state dependent means to study asymmetries in inflation drivers (inflation expectations, external prices, slack, ULCs...).

C1671: **Economic policy uncertainty spillovers in Europe and Greece**
Presenter: **Paraskevi Tzika**, University of Macedonia, Greece
Co-authors: Stylianos Fountas

Given the increasing uncertainty surrounding the European economies after the Eurozone crisis, an attempt is made to shed light on the spillovers of economic policy uncertainty among the European countries and the effects this uncertainty has on the Greek economy. At first, in order to identify and measure the spillovers of policy uncertainty in the Eurozone, a spillover index is constructed, based on a VAR model, for 7 Eurozone countries (from January 1998 until June 2019). The results show that the total connectedness among these countries' uncertainty is almost 50%. Moreover, France, Italy, Spain and the Netherlands are uncertainty exporters, while Greece, Germany and Ireland are uncertainty importers. Secondly, the case of Greece is investigated by constructing a VAR model and the respective generalized impulse response functions. The results indicate that both Greek and European uncertainty increase unemployment, while they reduce output and the Greek stock market index during the recent crisis period. For the pre-crisis period, the Greek economy seems to be affected more by the European uncertainty. Finally, there is positive interdependence between Greek and European policy uncertainty.

CG203 Room MAL 353 CONTRIBUTIONS IN HIGH-FREQUENCY ECONOMETRICS Chair: Markus Bibinger

C0553: **Likelihood evaluation through particle filter methods for high-frequency stochastic volatility models**
Presenter: **Antonio Santos**, University of Coimbra, Portugal

Volatility analysis is crucial for financial decision-making. High-frequency volatility analysis is feasible from intraday data availability, and through the application of stochastic volatility models to such data. New elements appeared as critical in the full characterization of high-frequency volatility evolution, and examples are the two-factor model and the presence of jumps in volatility. The variety of models increased, and using a Bayesian paradigm to estimate the models' parameters; direct likelihood evaluations are not feasible. The log-likelihood function value is essential to calculate the Bayes' factors, which is a measure used to choosing between competing models. A strategy is applied for evaluating the log-likelihood function of such complex models. The log-likelihood function evaluation makes use of techniques like importance sampling and particle filter methods. The results allow defining the best type of stochastic volatility model to use for volatility evolution analysis on two equities traded in US markets.

C1897: **Sources of global trading activity: A dynamic factor model for multivariate stock-market data**
Presenter: **Manuel Stapper**, WWU Muenster, Germany
Co-authors: Andreas Masuhr

Stock market trading activity serves as an instrument to measure the risk of assets. A dynamic factor model is used to describe the number of trades in time intervals in a multivariate setting. Assets from finance and pharmaceutical sector are considered jointly. Incorporating data from different stock exchanges around the globe narrows the window of missing information about news affecting the stock market. Besides idiosyncratic factors, the model allows for a market-wide common factor, two sector-specific factors and country-specific factors for each stock exchange. Parameters are estimated with a Particle Metropolis-Hastings procedure and the estimates are used to determine possible influences on trading activity for individual assets.

C1983: **The impact of periodicity on volatility-volume relations**
Presenter: **Yi Luo**, Lancaster University, United Kingdom
Co-authors: Zhen Wei

Opening, lunch, and closing of financial markets induce a periodic component in the volatility of high-frequency returns. However, the intraday volume and number of trades also display a prominent U curve that is still left to be investigated. We propose to use the Seasonal-Trend Decomposition Procedure Based on Loess to estimate the periodic component in volume and number of trades. We find that accounting for periodicity improves the explanatory power of both volume and number of trades on realized variance. Besides, the relationship between the average absolute return and volume (number of trades as well) can be better modeled using the mixture of two linear regression models during the trading day. With more analysis on the posterior probabilities of the mixing components, the average intraday volume and the number of trades display a higher effect on the absolute return in the morning relative to the rest of the day. The observed patterns indicate the need to decompose and analyze the periodicity not only in the realized variance but also in the volume and number of trades.

C1443: **Forecasting risk measures using intraday data in a generalized autoregressive score framework**
Presenter: **Xiaohan Xue**, ICMA Centre, University of Reading, United Kingdom
Co-authors: Emese Lazar

A new framework for the joint estimation and forecasting of dynamic Value-at-Risk (VaR) and Expected Shortfall (ES) is proposed by incorporating intraday information into a generalized autoregressive score (GAS) model to estimate risk measures in a quantile regression setup. We consider four intraday measures: the realized volatility at 5-min and 10-min sampling frequencies, and the overnight return incorporated into these two realized volatilities. In a forecasting study, the set of newly proposed semiparametric models is applied to 4 international stock market indices: the S&P 500, the Dow Jones Industrial Average, the NIKKEI 225 and the FTSE 100, and is compared with a range of parametric, nonparametric and semiparametric models including historical simulations, GARCH and the original GAS models. VaR and ES forecasts are backtested individually, and the joint loss function is used for comparisons. The results show that GAS models, enhanced with the realized volatility measures, outperform the benchmark models consistently across all indices and various probability levels.

CP001 Room Macmillan Hall and Crush Hall POSTER SESSION CFE Chair: Elena Fernandez Iglesias

C1633: **Efficient Bayesian computation for modeling dynamic counts**
Presenter: **Yuko Onishi**, Mitsubishi UFJ Morgan Stanley, Japan

Co-authors: Kaoru Irie, Shonosuke Sugasawa

Dynamic count data frequently appear in many scientific fields including finance, genomics, and social science. Although state space models based on the Poisson distribution are widely used, it is difficult to efficiently compute the posterior distribution due to the lack of conjugacy of normal distribution for the Poisson rate parameter even when the data is univariate. An efficient Bayesian computation method is proposed for multivariate Poisson state space models based on data augmentation. We use the negative binomial distribution as an approximation of the Poisson distribution and employ Polya-gamma data augmentation which enables us to compute the state variables efficiently. The proposed method is demonstrated through simulation and empirical studies.

C1726: **State space models for stochastic claims reserving**

Presenter: **Radek Hendrych**, Charles University, Czech Republic

The actuarial loss (claims) reserves represent estimates of insurers liabilities from all future claims obligations arising from insurance contracts currently in force and written in the past. There exist various actuarial (statistical) methods for calculation of outstanding claims reserves. The aim is to introduce and compare various methodological approaches to IBNR (incurred but not yet reported) claims reserving based on state space models and Kalman filter algorithms. In particular, if one transforms available claims data from the form of run-off triangles to the form of (multivariate) time series with missing observations, various state space models might be employed to project and/or interpolate IBNR claims reserves. Additionally, useful extensions of the loss reserving problem (e.g. dependent run-off triangles for correlated business lines or outliers in claims data) might be implemented when applying such a methodological framework. Results of numerical study for various claims data (univariate and multivariate ones) are presented.

C1821: **Modeling financial durations with limit order book information**

Presenter: **Tomoki Toyabe**, Keio University, Japan

Co-authors: Kentaro Asaba, Teruo Nakatsuma

It is a stylized fact that durations between executions in financial markets have intraday seasonality and autocorrelation. The Autoregressive Conditional Duration (ACD) model has been widely used to capture these characteristics. However, durations are also supposed to be affected by liquidity in the market. We propose a new ACD model that utilizes the limit order book information for reflecting the liquidity. In our empirical analysis, we applied the proposed ACD model to high-frequency stock price data in the Tokyo Stock Exchange and estimated it with an efficient Markov chain Monte Carlo method. We also conducted model comparison among different specifications of the proposed model.

C1826: **Bayesian modeling of high frequency stochastic volatility with intraday seasonality and skew heavy-tailed error**

Presenter: **Makoto Nakakita**, Keio University, Japan

Co-authors: Teruo Nakatsuma

Intraday high frequency data of asset returns exhibit not only typical characteristics (e.g., volatility clustering) but also a cyclical pattern of return volatility that is known as intraday seasonality. We extend the stochastic volatility (SV) model for application with such intraday high frequency data and develop an efficient Markov chain Monte Carlo sampling algorithm for Bayesian inference of the proposed model. Our modeling strategy is two-fold. First, we model the intraday seasonality in return volatility with orthogonal polynomials and estimate it along with the stochastic volatility simultaneously. Second, we incorporate a possibly skew and heavy-tailed error distribution into the SV model by assuming that the error distribution belongs to a family of generalized hyperbolic distributions such as variance-gamma, Student's t and their skew variants. As a demonstration of our new method, we estimate the proposed model with 1-minute and 5-minute return data of a stock index (TOPIX) and conduct comparison among competing model specifications with the widely applicable information criterion (WAIC). The results show that the SV model with skew variance-gamma error is the best in a volatile market.

C1913: **Machine learning solution of dynamic models with rational inattention**

Presenter: **Lukas Vacha**, Institute of Information Theory and Automation of the CAS, Czech Republic

Co-authors: Jozef Barunik

An asset pricing model is proposed under uncertainty with agents having quantile preferences and limited information processing capacity. In contrast to the standard asset pricing that relies on expected utility, we introduce a dynamic quantile model for asset pricing, in which the agent maximizes stream of future quantile utilities. The agent has a limited amount of attention since the information she obtains is costly. In our model, the agent maximizes stream of her future quantile utilities according to her quantile utility preferences subject to information costs constraints. We solve this high-dimensional problem using machine learning tools leading to a sequence of supervised learning problems. This approach makes this task computationally feasible. Results indicate that there is a significant benefit when a standard expected utility is expanded into quantile preference utilities.

C1494: **Financial distress risk and stock price crashes**

Presenter: **Panayiotis Andreou**, Cyprus University of Technology, Cyprus

Co-authors: Neophytos Lambertides, Christoforos Andreou

A strong positive relationship between changes in firms distress risk and future stock price crashes is studied. Particularly, changes in distress risk can predict stock price crashes as far as three months ahead in the future. The results show that the crash-distress relationship is more pronounced when firms information asymmetry is higher, as captured by firms accounting opacity and stock liquidity. Interestingly, the findings support that the impact of distress risk changes on future stock price crashes is stronger: i) during investor sentiment-correction periods, and ii) periods with heightened market default risk. The finding are of interest to investors who wish to take long-run positions in the stock market because stock price crash risk cannot be easily diversified away. In this vein, investors should be cautious of a firms distress risk as short-term increases could be an early warning sign for forthcoming crash risk problems.

C2038: **Analysis of the cryptocurrencies market evolution using networks**

Presenter: **Pilar Grau**, Universidad Rey Juan Carlos, Spain

Co-authors: Luis Miguel Doncel

Cryptocurrencies have begun to transform the global financial system. Both their number and trading volume have increased in the recent years. The purpose of this research is to analyze the structure of the cryptocurrency market through the correlations of 119 of its constituent in five different years using methods of random matrix theory and minimum spanning trees. We use network analysis to represent the correlations of the cryptocurrency market and detecting distinct community structures in its minimum spanning tree. Using centrality analysis with betweenness, closeness, degree, eigenvector centrality, and PageRank we demonstrate the importance of cryptocurrencies and their evolution over time. The results can be useful when constructing cryptocurrency investment portfolios.

Sunday 15.12.2019 16:35 - 18:15 Parallel Session J – CFE-CMStatistics

EI016 Room Beveridge Hall MEASUREMENT ERROR MODELS AND BEYOND Chair: Wenqing He

E0174: **Estimation in GEE models for clustered/longitudinal data with covariate measurement error**
Presenter: **Jeff Buzas**, The University of Vermont, United States

Estimation and inference in cluster/longitudinal population averaged models is considered in the presence of covariate measurement error, with emphasis on logistic regression. A general approach to bias correction, resulting in approximately consistent estimators, is considered when auxiliary data in the form of replicates or instrumental variables are available. For the logistic model, the approach yields unbiased estimating equations resulting in fully consistent estimators.

E1386: **Progressive multi-state models with misclassified states**
Presenter: **Grace Yi**, University of Western Ontario, Canada

Progressive multi-state models are commonly used in studies of disease progression. Methods developed under this framework, however, are often challenged by misclassification in states. We will discuss issues concerning continuous time progressive multi-state models with state misclassification. We will describe inference methods using both the likelihood and pairwise likelihood methods that are based on joint modelling of the transition and misclassification processes. The performance of estimation procedures is evaluated by numerical studies.

E0173: **Variable selection and estimation in generalized linear models with measurement error**
Presenter: **Liqun Wang**, University of Manitoba, Canada

The variable selection and estimation problems in linear and generalized linear models are studied when some of the predictors are measured with error. We demonstrate how measurement error (ME) affects the selection results and propose regularized instrumental variable (RIV) methods to correct for the ME effects. We show that the proposed methods have the oracle property in a linear model and we derive their asymptotic distribution under general conditions. We also investigate the performances of the methods in generalized linear models. Our simulation studies show that the RIV methods outperform the naive method in both linear and some generalized linear models. Finally, the proposed method is applied to a real dataset.

EO711 Room CLO B01 FUNCTIONAL AND SHAPE DATA ANALYSIS Chair: Anuj Srivastava

E0447: **Smoothing splines on Riemannian manifolds, with applications to 3D shape space**
Presenter: **Ian Dryden**, University of Nottingham, United Kingdom
Co-authors: Kwang-Rae Kim, Huiling Le

There has been increasing interest in statistical analysis of data lying in manifolds. A smoothing spline fitting method is generalized to Riemannian manifold data based on the technique of unrolling and unwrapping originally proposed by Jupp and Kent for spherical data. In particular a fitting procedure is developed for shapes of configurations in general m-dimensional Euclidean space, extending previous work for two dimensional shapes. It is shown that parallel transport along a geodesic on Kendall shape space is linked to the solution of a homogeneous first-order differential equation, some of whose coefficients are implicitly defined functions. This finding enables one to approximate the procedure of unrolling and unwrapping by simultaneously solving such equations numerically, and so to find numerical solutions for smoothing splines fitted to higher dimensional shape data. This fitting method is applied to the analysis of simulated 3D shapes and to some dynamic 3D peptide data.

E0452: **Curvature correction for distance-based learning on shape space**
Presenter: **Philipp Harms**, University of Freiburg, Germany

As statistics on manifolds can be prohibitively slow, linearizations are often used in practice to obtain approximations at reasonable computational cost. These approximations are, however, distorted by the curvature of the space. We introduce curvature corrections for distance-based algorithms such as multi-dimensional scaling or agglomerative clustering and show some applications to shape analysis on landmark spaces.

E1014: **The use of landmarks within elastic planar shape analysis**
Presenter: **Justin Strait**, University of Georgia, United States

In shape analysis, landmarks are generally identified as important features of the shape which are mathematically and/or anatomically relevant. While recent advances in the field have been focused on shape representations which treat the underlying contour as an object in infinite-dimensional space, we will discuss some uses of landmarks as possibly useful inferential tools in this context. In the unsupervised learning scenario, landmarks can be thought of as a low-dimensional set of points along shape contours which "approximate" the underlying object well. While other methods may produce estimates of underlying features, the advantage of the model-based method we propose is the ease of interpretability in identifying landmarks as these important features. If class labels are also known in addition, we can treat landmarks as latent variables, in the sense that we can identify them as important points which discriminate between shape classes optimally. In both settings, we propose a hierarchical, model-based approach, producing estimates of uncertainty for landmark locations. We will discuss modeling issues which arise from the Bayesian perspective, and demonstrate the use of these models on both simulated and real data.

E0480: **Functional PCA on shape space for gait analysis and assessment**
Presenter: **Nadia Hosni**, University of Lille, France
Co-authors: Boulbaba Ben Amor, Hassen Drira, Faten Chaieb

The functional Principal Component Analysis (fPCA) on the Kendall shape space is used to study 3D human shape trajectories. Various problems, including – gait recognition, gait classification (normal vs. pathological), gender classification and physical performance assessment from 3D skeletal data – are addressed. The key idea is to transform initial high-dimensional shape trajectories to a compact set of uncorrelated variables. That is, the proposed Kendall fPCA allows representation of their variation around the mean trajectory in a lower dimensional submanifold, in terms of principal modes of variation. Acquired using conventional IR MoCap senors (e.g. Vicon) or cost-effective depth cameras (e.g. Kinect), sequences of skeletal data are first mapped to the shape space of 3D landmark configurations and viewed as time-parameterized shape trajectories. The main barrier to apply fPCA is the non-linear structure of the space of interest. We accommodate fPCA formulation to account for the geometric structure of the Kendall shape space. The elastic metric and the geometric tools previously defined allowed us to align temporally the trajectories and to approximate the Frechet mean trajectory. Kendall fPCA is carried out by log-map the original trajectories to tangent spaces around the Frechet mean, and then performed a classical fPCA on the vector fields lying to the linear tangent spaces of log-mapped data.

EO697 Room MAL B02 STATISTICAL METHODS FOR RISK MANAGEMENT Chair: Marie Kratz

E0571: **Tail risk and style dependence in the fund industry: A multivariate extreme value approach**
Presenter: **Julien Hambuckers**, University of Liege - HEC Liege, Belgium
Co-authors: Linda Mhalla, Marie Lambert

With the recent financial crisis and the increasing interconnectedness of financial institutions, regulators have started to pay more attention to the concept of systemic risk. Whereas most of the systemic risk literature focuses on the banking industry, there has also been a growing interest in the contribution of unregulated funds, e.g. hedge funds. We propose to look at systemic risk in a universe of funds by the prism of conditional extremal

tail dependence across the different investment strategies of the funds. Relying on univariate and multivariate extreme value theory, we first model the dynamics of style-specific extreme fund losses with a non-stationary generalized Pareto distribution depending on the return characteristics and then study the dynamics of the tail dependencies between fund styles conditional on measures of the economic uncertainty and the stock market performance. We show that economic uncertainty and market stress influence the links between strategy-specific extreme losses.

E0627: **Extremes of extendible random vectors**
Presenter: **Johanna Neslehova**, McGill University, Canada
Co-authors: Klaus Herrmann, Marius Hofert

Classical extreme value theory is concerned with the limiting behavior of maxima of independent and identically distributed random variables under appropriate location-scale transformations. When working with large portfolios, the assumption of independence may no longer be appropriate. We will explore the weak limits of maxima of identically distributed random variables which are neither independent nor form a locally dependent time series. A particularly tractable case is that of an extendible sequence of random variables whose dependence is Archimedean. As we will see, the possible limits are no longer extreme-value distributions, but an asymptotic theory for maxima can nonetheless be developed and is driven by the properties of the Archimedean generator. Extensions of these findings to other extendible sequences of random variables will also be discussed.

E0787: **Risk aggregation and the effect of reinsurance**
Presenter: **Alexandra Dias**, University of York, United Kingdom

Insurance companies must fulfil solvency capital requirements in order to ensure that they can meet their future obligations to policyholders. The solvency capital requirement is a risk management tool essential when extreme catastrophic events happen resulting in large possibly interdependent claims. The purpose is to study empirically the problem of aggregating the risks coming from several insurance lines of business and analyses the effect of reinsurance in the level of aggregate risk. Our starting point is to use a hierarchical risk aggregation model, which was initially based on 2-dimensional elliptical copulas. We use copulas from the Archimedean family. The results show that a mixture of copulas can provide a better fit to the data than the plain (single) copulas and consequently avoid overestimation or underestimation of the capital requirement of an insurance company. We also investigate the significance of reinsurance in reducing the insurance company's business risk and its effect on diversification. Our empirical study reveals that reinsurance might reduce the effect of diversification for insurance companies with multiple business lines.

E1097: **(F)CLT for functionals of quantile and dispersion estimators: Applications in risk management**
Presenter: **Marcel Brautigam**, ESSEC Business School & Sorbonne University, France
Co-authors: Marie Kratz

CLTs and FCLTs for Functionals of Quantile and Dispersion Estimators are provided. Those will allow us to quantify and explain the statistical pro-cyclicality of standard risk measures. First, we derive the joint bivariate asymptotic distributions of functions of quantile estimators (the non-parametric sample quantile and the parametric location-scale quantile estimator) with functions of measure of dispersion estimators (the sample variance, sample mean absolute deviation or any higher order absolute central sample moment) - assuming an underlying identically and independently distributed sample. In a second step, we extend the joint asymptotics between the sample quantile and any absolute centred sample moment for a broad class of augmented GARCH processes. These results support the empirical findings about the pro-cyclicality of traditional risk measurements. In particular, it proves that part of the procyclicality is intrinsically caused by the way of the historical risk estimation. Further, the exact degree of pro-cyclicality depends on the choice of the risk measure as well as the measure of dispersion considered but does not vanish in any case.

EO542 Room MAL B04 TOPICS ON DIMENSION REDUCTION AND KERNEL METHODS Chair: Andreas Artemiou

E0388: **A weighted learning approach for sufficient dimension reduction in binary classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South

Since the proposal of the sliced inverse regression (SIR), inverse-regression methods have been widely used for sufficient dimension reduction (SDR). In binary classification, the inverse-regression methods including SIR often suffer from the lack of resolution of the binary response. We propose a weighted large-margin classifiers to recover the central subspace. Toward this, we establish that the gradient vector of the weighted large-margin classifier is unbiased for SDR if the corresponding weighted loss function is Fisher consistent. This enables us to propose what we call weighted outer-product of gradients (wOPG) method for SDR in binary classification. The proposed wOPG method can recover the central subspace exhaustively without linearity condition or constant variance condition and shows promising performance for both simulated and real data examples.

E0396: **Central quantile subspace**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States

Quantile regression (QR) is becoming increasingly popular due to its relevance in many scientific investigations. There is a great amount of work about linear and nonlinear QR models. Specifically, nonparametric estimation of the conditional quantiles received particular attention, due to its model flexibility. However, nonparametric QR techniques are limited in the number of covariates. Dimension reduction offers a solution to this problem by considering low-dimensional smoothing without specifying any parametric or nonparametric regression relation. Existing dimension reduction techniques focus on the entire conditional distribution. We, on the other hand, turn our attention to dimension reduction techniques for conditional quantiles and introduce a new method for reducing the dimension of the predictor X. The novelty is threefold. We start by considering a single index quantile regression model, which assumes that the conditional quantile depends on X through a single linear combination of the predictors, then extend to a multi index quantile regression model, and finally, generalize the proposed methodology to any statistical functional of the conditional distribution. The results suggest that this method has a good finite sample performance and often outperforms existing methods.

E0640: **On sufficient dimension reduction for functional data**
Presenter: **Jun Song**, University of North Carolina at Charlotte, United States
Co-authors: Bing Li

A general theory and estimation methods for functional linear sufficient dimension reduction are developed, where both the predictor and the response can be random functions, or even vectors of functions. Unlike the existing dimension reduction methods, our approach does not rely on the estimation of conditional mean and conditional variance. Instead, it is based on a new statistical construction –the weak conditional expectation, which is based on Carleman operators and their inducing functions. Weak conditional expectation is a generalization of conditional expectation. Its key advantage is to replace the projection on to an L2-space – which defines conditional expectation – by projection on to an arbitrary Hilbert space, while still maintaining the unbiasedness of the related dimension reduction methods. This flexibility is particularly important for functional data, because attempting to estimate a full-fledged conditional mean or conditional variance by slicing or smoothing over the space of vector-valued functions may be inefficient due to the curse of dimensionality. We evaluated the performances of our new methods by simulation and in several applied settings.

E1995: **The predictive power of kernel principal components regression**
Presenter: **Ben Jones**, Cardiff University, United Kingdom
Co-authors: Andreas Artemiou, Bing Li

A well-known empirical phenomenon in statistics is that the higher-ranking principal components of a predictor variable tend to have greater

squared correlations with a response variable than the lower-ranking ones, even though the extraction procedure is unsupervised. This was originally observed in the classical setting, where it is assumed that a linear model relates the response with the predictor. In this setting, theoretical analyses have proven that, under a uniformity assumption on the regression coefficients or the covariance matrix of the predictor, this tendency holds at population-level. Further studies have established this result in more general settings, including in single-index and conditional independence models. The principal components procedure can be extended, using the kernel trick, to nonlinear directions in the data. The predictive tendency, using the nonlinear components, has also been empirically recognised in this setting. Recent research is detailed which establishes this tendency, in this nonlinear setting, at population-level. The first framework is that of nonparametric regression. The second, much more general, framework is that where the response conditional on the predictor has an arbitrary or random distribution.

EO600 Room MAL B18 METHODS FOR MISSING DATA Chair: Anders Lundquist

E0766: **Latent class response propensity models for non-ignorable item nonresponse in surveys**
Presenter: **Jouni Kuha**, London School of Economics, United Kingdom

When missing data are produced by a non-ignorable non-response mechanism, analysis of the observed data should include a model for the probabilities of responding. We propose such models for non-response in survey questions which are treated as measures of latent constructs and analysed by using latent variable models. The non-response models that we describe include additional latent variables (latent response propensities) which determine the response probabilities. We argue that this model should be specified as flexibly as possible, and we propose models where the response propensity is a categorical variable (a latent response class). This can be combined with any latent variable model for the survey items, and an association between the latent variables measured by the items and the latent response propensities then implies a model with non-ignorable non-response. We consider in particular such models for the analysis of data from cross-national surveys, where the non-response model may also vary across the countries. The models are applied to data on welfare attitudes in 29 countries in the European Social Survey.

E0952: **Estimating the population partly conditional mean using longitudinal cohort data with non-ignorable drop-out**
Presenter: **Maria Josefsson**, Centre for Demographic and Ageing Research, Sweden

Understanding how cognition changes during normal aging is important, since these more subtle changes still may affect day-to-day function, as well as for differentiating between normal and pathological states. Studies of cognitive aging using longitudinal data often result in highly selected samples due to selective study enrollment and attrition. An additional methodological challenge is practice effects, resulting in improved or maintained test scores despite a cognitive decline. These challenges may bias study finding and severely distort the ability to generalize to the target population even in well-designed studies. We propose an approach for estimating the finite population average of a longitudinal continuous cognitive outcome conditioning on being alive at a specific age, i.e. the population partly conditional mean. Specifically, we develop a flexible Bayesian semi-parametric predictive estimator, when longitudinal auxiliary information is known for all units in the target population. By specifying priors for the sensitivity parameters our approach allows uncertainty about untestable assumptions. The proposed approach is motivated by 15-year longitudinal data from the Betula longitudinal cohort study. We apply our approach to study normal cognitive aging with the aim to generalize findings to the target population.

E1254: **Missing data: A unified taxonomy guided by conditional independence**
Presenter: **Marco Doretti**, University of Perugia, Italy
Co-authors: Sara Geneletti, Elena Stanghellini

Recent work attempts to clarify the not always well-understood difference between realised and everywhere definitions of missing at random (MAR) and missing completely at random. Another branch of the literature exploits always-observed covariates to give variable-based definitions of MAR and missing completely at random. We develop a unified taxonomy encompassing all approaches. In this taxonomy, the new concept of complementary MAR is introduced, and its relationship with the concept of data observed at random is discussed. All relationships among these definitions are analysed and represented graphically. Conditional independence, both at the random variable and at the event level, is the formal language we adopt to connect all these definitions. Both the univariate and the multivariate case are covered. Attention is paid to monotone missingness and to the concept of sequential MAR. Specifically, for monotone missingness, we propose a sequential MAR definition that might be more appropriate than both everywhere and variable-based MAR to model dropout in certain contexts.

E1279: **Bayesian nonparametric methods for longitudinal outcomes missing not at random**
Presenter: **Antonio Linero**, Florida State University, United States
Co-authors: Michael Daniels

The setting of a longitudinal outcome subject to nonignorable missingness is considered. This requires the specification $f(y,r)$ for the joint model of the response y and missing data indicators r. We argue that the obvious Bayesian nonparametric approaches to joint modeling which have been applied in the literature run afoul of the inherent identifiability issues with nonignorable missingness, leading to posteriors with dubious theoretical behavior and producing questionable inferences. As an alternative, we propose an indirect specification of a prior on the observed data generating mechanism $f(y_{obs},r)$, which is fully identified given the data. This prior is then used in conjunction with an identifying restriction to conduct inference. Advantages of this approach include a flexible modeling framework, access to simple computational methods, flexibility in the choice of "anchoring" assumptions, strong theoretical support, straightforward sensitivity analysis, and applicability to non-monotone missingness.

EO677 Room MAL B20 CAUSAL INFERENCE USING OBSERVATIONAL LONGITUDINAL DATA Chair: Shaun Seaman

E1455: **Robust Q-learning**
Presenter: **Ashkan Ertefaie**, University of Rochester, United States
Co-authors: Robert Strawderman

Q-learning is a regression-based approach that is widely used to formalize the development of an optimal dynamic treatment strategy. Finite dimensional working models are typically used to estimate certain nuisance parameters, and misspecification of these working models can result in residual confounding and/or efficiency loss. We propose a robust Q-learning approach which allows estimating such nuisance parameters using data-adaptive techniques. We study the asymptotic behavior of our estimators and provide simulation studies that highlight the need for and usefulness of the proposed method in practice. We use the data from the Extending Treatment Effectiveness of Naltrexone multistage randomized trial to illustrate our proposed methods.

E1427: **A Bayesian factor analysis model for evaluating an intervention using observational panel data on multiple outcomes**
Presenter: **Pantelis Samartsidis**, University of Cambridge, United Kingdom
Co-authors: Shaun Seaman, Silvia Montagna, Andre Charlett, Matthew Hickman, Daniela de Angelis

A problem frequently encountered in many areas of scientific research is that of estimating the impact of a non-randomised binary intervention on an outcome of interest using time-series data on units that received the intervention (treated) and units that did not (controls). One popular estimation method in this setting is based on the factor analysis (FA) model. The FA model is fitted to the pre-intervention outcome data on treated units and all the outcome data on control units, and the counterfactual treatment-free post-intervention outcomes of the former are predicted from the fitted model. Intervention effects are estimated as the observed outcomes minus these predicted counterfactual outcomes. We propose two extensions of the FA model for estimating intervention effects: 1) the joint modelling of multiple outcomes to exploit shared variability, and 2) an

autoregressive structure on factors to account for temporal correlations in the outcome. Using simulation studies, we show that both extensions can improve the precision of the intervention effect estimates: the first when the number of pre-intervention measurements is small; the second when the number of control units is small. We apply our method to estimate the impact of stricter alcohollicensing policies on alcohol-related harms.

E0990: **Using sequential trials to estimate treatment effects in longitudinal observational data**
Presenter: **Ruth Keogh**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: Shaun Seaman, Jon Michael Gran, Stijn Vansteelandt

Randomized controlled trials are the gold standard for estimating causal effects of treatments on health outcomes, but can be infeasible or unethical. Longitudinal observational data offer the possibility of estimating treatment effects over long periods of follow-up and in diverse populations. However, to do this we must tackle the challenge of time-dependent confounding. Several methods have been described for estimating causal treatment effects on survival using longitudinal observational data. The focus is on the sequential trials approach, which involves creation of a sequence of artificial trials from new time origins within an observational cohort. The analysis uses pooled Cox regression and time-dependent confounding is addressed through baseline covariate adjustment, censoring people upon deviation from their baseline treatment group, and inverse probability of censoring weighting. The sequential trials approach, despite being intuitive and straightforward to implement, has not previously been compared with alternative methods, either empirically or in terms of theoretical properties. We will use the potential outcomes framework to explain what is being estimated in the sequential trials approach, and contrast this with other methods. We will also show how sequential trials can be used to compare survival probabilities if different treatment regimes were applied in the target population.

E0809: **Handling time-dependent confounding using a structural nested cumulative survival time model**
Presenter: **Shaun Seaman**, University of Cambridge, United Kingdom
Co-authors: Shaun Seaman, Oliver Dukes, Ruth Keogh, Stijn Vansteelandt

Observational studies that attempt to assess the effect of a time-varying exposure on a survival outcome typically suffer from time-varying confounding bias. Marginal structural models, fitted by inverse probability weighting, can be used, but these can be subject to the problem of highly variable weights when the confounders are strongly predictive of the exposure or when the exposure is continuous. Structural nested accelerated failure time models can be fitted by g-estimation, but this requires artificial recensoring, which causes loss of information. We have developed an alternative method using a structural nested cumulative survival time model (SNCSTM). This method avoids inverse probability weighting and artificial recensoring. It also allows investigation of effect modification by time-dependent variables. The SNCSTM assumes that intervening to set exposure at time t to zero has an additive effect on the subsequent conditional hazard given exposure and confounder histories when all subsequent exposures have already been set to zero. We show how to estimate the exposure effect using standard software for generalised linear models and describe a more efficient estimator that is available in closed form. We apply our methods to estimate the effect of delaying initiation of treatment with DNase on survival in patients with Cystic Fibrosis.

EO755 Room MAL B35 TRADITIONAL AND MODERN TIME SERIES MODELS Chair: Pauliina Ilmonen

E0522: **On modelling and estimation of stationary processes**
Presenter: **Marko Voutilainen**, Aalto University, Finland

Stationary processes form an important class of stochastic processes that has been extensively studied in the literature. Their applications include modelling and forecasting numerous real-life phenomenon including natural disasters, sustainable energy sources, sales and market movements. We present a novel way for modelling and estimating n-dimensional strictly stationary processes, both in discrete and continuous time. The approach is based on the observation that stationary processes are characterized by an AR(1) type of (matrix) equation in discrete time, and by n-dimensional Langevin equation in continuous time. As a consequence, we obtain a continuous time algebraic Riccati equation for the model parameter matrix given by the characterization. The Riccati equation provides us with a natural estimator of the model parameter that inherits consistency and asymptotic properties from the autocovariance function of the stationary process.

E0547: **Dimension reduction for time series in a blind source separation context**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Klaus Nordhausen, Markus Matilainen, Jari Miettinen, Joni Virta

Multivariate time series observations are increasingly common in many fields of science but the complex dependencies between multiple time series often yield to intractable models with large number of parameters. An alternative approach is to first reduce the dimension of the series and then model the resulting uncorrelated univariate time series. Blind source separation (BSS) offers a popular and effective framework for this. We review some dimension reduction tools for time series in a BSS context. In specific, we propose an estimator which is developed for identifying components which exhibit volatility clustering. The theoretical properties of the estimator are discussed and an example is provided to illustrate the method.

E0602: **On adaptive functional data depths**
Presenter: **Sami Helander**, Aalto University School of Science, Finland
Co-authors: Stanislav Nagy, Germain Van Bever, Lauri Viitasaari, Pauliina Ilmonen

Typically, in the functional context, data depth approaches heavily emphasize the location of the functions in the distribution, therefore often missing important shape or roughness features. Commonly, these depth approaches either integrate pointwise depth values to achieve a global value, or measure the expected distance from a function to the distribution. We introduce a new class of functional depths, based on the distribution of depth values along the domain, and discuss their properties. We study the asymptotic properties of these Jth order kth moment integrated depths, and illustrate their usefulness in supervised functional classification. In particular, we demonstrate the importance of receptivity to shape variations, and show that, similarly to existing depth notions, the new class of depth functions takes into account the variation in location, while remaining receptive to variations in shape and roughness as well.

E0637: **Pretty predictable models**
Presenter: **Tommi Sottinen**, University of Vaasa, Finland

A class of stochastic processes that are generated by a so-called invertible Gaussian Volterra process is considered. By this we mean that we can, in an adaptive way, recover an underlying Brownian motion that generates the same filtration as the stochastic process under consideration. By using the underlying Brownian motion, we construct explicitly the regular future conditional law of our stochastic process conditioned on the past. Examples include fractional Brownian motions and geometric fractional Brownian motions.

EO703 Room MAL B36 NEW DIRECTIONS IN STATISTICAL LEARNING Chair: Sunyoung Shin

E0988: **Assessment of case influence in support vector machine**
Presenter: **Yoonkyung Lee**, Ohio State University, United States
Co-authors: Shanshan Tu, Yunzhang Zhu

Support vector machine (SVM) is a very popular technique for classification. A key property of SVM is that its discriminant function depends only on a subset of data points called support vectors. This comes from the representation of the discriminant function as a linear combination of kernel functions associated with individual cases. Despite the direct relation between each case and the corresponding coefficient in the representation, the

influence of cases and outliers on the classification rule has not been examined formally. Borrowing ideas from regression diagnostics, we define case influence measures for SVM and study how the classification rule changes as each case is perturbed. To measure case sensitivity, we introduce a weight parameter for each case and reduce the weight from one to zero to link the full data solution to the leave-one-out solution. We develop an efficient algorithm to generate case-weight adjusted solution paths for SVM. The solution paths and the resulting case influence graphs facilitate evaluation of the influence measures and allow us to examine the relation between the coefficients of individual cases in SVM and their influences comprehensively. We present numerical results to illustrate the benefit of this approach.

E1282: **Analyzing group fMRI with multilayer network embedding methods**
Presenter: **James Wilson**, University of San Francisco, United States

Learning interpretable features from complex multilayer networks is a challenging and important problem. The need for such representations is particularly evident in multilayer networks of the brain, where nodal characteristics may help model and differentiate regions of the brain according to individual, cognitive task, or disease. Motivated by this problem, we introduce the multi-node2vec algorithm, an efficient and scalable feature engineering method that automatically learns continuous node feature representations from multilayer networks. A second-order random walk sampling procedure that efficiently explores the inner- and intra- layer ties of the observed multilayer network is utilized to identify multilayer neighborhoods. Maximum likelihood estimators of the nodal features are identified through the use of the Skip-gram neural network model on the collection of sampled neighborhoods. We demonstrate the efficacy of multi-node2vec on a multilayer functional brain network from resting state fMRI scans over a group of 74 healthy individuals and 70 patients with varying degrees of schizophrenia. Findings reveal that multi-node2vec identifies regional characteristics that closely associate with the functional organization of the brain and offer insights into the differences between the patient and healthy groups.

E1421: **Composite hybrid basis approach incorporating residual connectivity in task fMRI data**
Presenter: **Michelle Miranda**, University of Victoria, Canada
Co-authors: Jeff Morris

A composite hybrid basis approach is proposed to model brain spatial and temporal correlation in task fMRI data. Our model has a better detection power than the state-of-the-art methods. The greater power is obtained by borrowing both spatial and temporal information through the carefully designed basis. First, the composite hybrid basis provides a sparse spatial representation of the brain by accounting for local (within ROIs) and distant (between ROIs) correlations while yielding full Bayesian inference at the voxel and ROI level with incredible computational speed. Second, the model allows for free full Bayesian inference on the residual connectivity, which can help scientist gain insights into the underlying brain function. We apply our model to the Example Subject of the Working Memory Task of the Human Connectome Project.

E1422: **Functional regression analysis of distributional data using quantile functions**
Presenter: **Hojin Yang**, University of Nevada Reno, United States
Co-authors: Veerabhadran Baladandayuthapani, Arvind Rao, Jeff Morris

The aims are to look at the subject specific distribution from observing the large number of repeated measurements for each subject and to determine how a set of covariates effects various aspects of the underlying subject-specific distribution, including the mean, median, variance, skewness, heavy tailedness, and various upper and lower quantiles. To address these, we develop a quantile functional regression modeling framework that models the distribution of a set of common repeated observations from a subject through the quantile function. To account for smoothness in the quantile functions, we introduce novel basis functions adapting to the features of a given data set. Then, we build a Bayesian framework that uses nonlinear shrinkage of basis coefficients to regularize the functional regression coefficients and allows fully Bayesian inferences after fitting a Markov chain Monte Carlo. We demonstrate the benefit of the basis space modeling through simulation studies, and illustrate the method using a biomedical imaging data set in which we relate the distribution of pixel intensities from a tumor image to various demographic, clinical, and genetic characteristics.

EO795 Room MAL G13 LATENT VARIABLE MODELS FOR COMPLEX DATA Chair: Silvia Cagnone

E0760: **Dealing with uncertainty in automated test assembly problems**
Presenter: **Giada Spaccapanico Proietti**, University of Bologna, Italy
Co-authors: Stefania Mignani, Mariagiulia Matteucci

Automated test assembly (ATA) models are intended to build standardized parallel test forms starting from an item bank. A general framework for ATA consists in adopting linear models which are solved by commercial solvers. Those solvers are not always able to find solutions for highly constrained and large-sized ATA instances. Moreover, all parameters are assumed to be fixed and known, a hypothesis that is not true for estimates of item response theory parameters. These restrictions motivated us to find an alternative way to define and solve ATA models. First, we suggest a chance-constrained approach, which allows maximizing the α-quantile of the empirical distribution function of the test information function obtained by bootstrapping the calibration process. Secondly, we adapt a stochastic meta-heuristic called simulated annealing for solving the ATA models. This technique can handle large-scale models and non-linear functions of which the chance constraints are an example and avoids local optima. A Lagrangian relaxation helps to find the most feasible/optimal solution. Several simulations are performed and the solutions are compared to the results of CPLEX 12.8.0 Optimizer. The algorithms are coded in the open-source programming language Julia.

E0841: **On the role of latent variables in characterizing some skew distributions**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia

The use of latent variables is considered to provide skew extensions of the multivariate normal and t densities. Various models are contrasted with the focus on the so-called canonical fundamental skew normal and skew t distributions. A critical comparison of the various models is provided in their use as component distributions in mixture models for the clustering of several data sets.

E0873: **Dimension-wise likelihood estimation of latent vector autoregressive models**
Presenter: **Silvia Bianconcini**, University of Bologna, Italy
Co-authors: Silvia Cagnone

Approximate methods are considered for likelihood inference to longitudinal and multidimensional data within the context of health science studies. The complexity of these data necessitates the use of sophisticated statistical models, that can pose significant challenges for model fitting in terms of computational speed, memory storage, and accuracy of the estimates. The methodology is motivated by a study that examines the temporal evolution of the mental status of the US elderly population between 2006 and 2010. We propose modeling the individual mental status as a latent process also accounting for the effects of individual specific characteristics, such as gender, age, and years of educational attainment. We describe the specification of a latent vector autoregressive model within the generalized linear latent variable framework, and its efficient estimation using a recent technique, called dimension-wise quadrature.

E2003: **Mixtures of seemingly unrelated linear regression models**
Presenter: **Gabriele Soffritti**, University of Bologna, Italy
Co-authors: Giuliano Galimberti

Finite mixtures of Gaussian linear regression models represent a flexible tool to perform linear regression analysis in the presence of a finite number of heterogeneous populations, each of which is characterized by a different Gaussian linear regression model. These models naturally arise when

relevant categorical predictors are omitted from a regression model. With several responses, such an approach makes it possible to take into account the correlation among responses that typically occur in longitudinal data, time-series data or repeated measures. In most of the models developed so far, the same regressors have to be used for all responses. This restriction is relaxed by allowing different regressors for each response, as in the seemingly unrelated regression framework. Parsimonious models are specified, by constraining the component-covariance matrices using a parameterisation that exploits their spectral decomposition. Details about model identification and maximum likelihood estimation are given. The usefulness of these models is shown through the analysis of a real dataset. The consistency of the maximum likelihood estimator under the proposed models is proved. The behaviour of this estimator in the presence of finite samples is numerically evaluated through the analysis of simulated datasets.

EO636 Room MAL G14 BAYESIAN INFERENCE AND COMPUTATIONAL ADVANCES FOR LARGE DATA Chair: Andee Kaplan

E0288: **Bayesian factor analysis for inference on interactions**
Presenter: **Federico Ferrari**, Duke University, United States
Co-authors: David Dunson

The motivation comes from the problem of inference on interactions among chemical exposures impacting human health outcomes. Chemicals often co-occur in the environment or in synthetic mixtures and as a result exposure levels can be highly correlated. We propose a latent factor joint model, which includes shared factors in both the predictor and response components while assuming conditional independence. By including a quadratic regression in the latent variables in the response component, we induce flexible dimension reduction in characterizing main effects and interactions. We propose a Bayesian approach to inference under this Factor analysis for INteractions (FIN) framework. Through appropriate modifications of the factor modeling structure, FIN can accommodate higher order interactions and multivariate outcomes. We provide theory on posterior consistency and the impact of misspecifying the number of factors. We evaluate the performance using a simulation study and data from the National Health and Nutrition Examination Survey (NHANES). Code is available on GitHub.

E0914: **Posterior prototyping for Bayesian entity resolution**
Presenter: **Andee Kaplan**, Colorado State University, United States
Co-authors: Brenda Betancourt, Rebecca Steorts

Entity resolution (record linkage or de-deduplication) is the process of merging noisy databases to remove duplicate entities, often in the absence of a unique identifier. One major challenge of linked data is identifying the most representative record to pass to an inferential or predictive task - the downstream task. To bridge the gap between entity resolution and the downstream task, we propose four methods - prototyping - to choose a representative record from linked data. The result is a representative data set to be passed on to the downstream task. To illustrate our proposed methodology, we first perform Bayesian entity resolution where the error can be propagated through to the downstream task. Second, we evaluate our proposed methods for prototyping. Third, we consider the downstream task of linear regression. The proposed methodology is illustrated and evaluated on five entity resolution data sets.

E0944: **Non-reversible parallel tempering: An embarrassingly parallel MCMC scheme**
Presenter: **Saifuddin Syed**, University of British Columbia, Canada
Co-authors: Alexandre Bouchard, George Deligiannidis, Arnaud Doucet

MCMC methods are widely used to approximate intractable expectations with respect to high-dimensional un-normalized probability distributions. We construct a Markov chain with the desired stationary distribution to explore our sample space. In theory, the chain should accurately explore the state space asymptotically, but in practice, it can get trapped exploring local regions of high probability and suffer from poor mixing in a finite time. Parallel tempering (PT) algorithms were introduced to tackle this issue. We delegate the task of exploration to additional heated chains running in parallel with better mixing properties. They then communicate with the target chain of interest and help it discover new unexplored regions of the sample space. Since their introduction in the 90's, PT algorithms are still extensively used to improve mixing in hard sampling problems arising in statistics, physics, computational chemistry, phylogenetics, and machine learning. We will give an introduction to PT algorithms, determine their scaling behaviour, efficiency, and limitations. Consequentially, we will establish the theoretically optimal non-reversible version of PT for a general class of sampling problems, as well an efficient and easy to implement adaptive algorithm.

E1387: **Bayesian modeling for large spatio-temporal data: An application to mobile networks**
Presenter: **Annalisa Cadonna**, WU, Vienna University of Economics and Business, Austria
Co-authors: Andrea Cremaschi, Alessandra Guglielmi, Fernando Quintana

Spatio-temporal areal data can be seen as a collection of time series which are spatially correlated, according to a specific neighboring structure. We propose a hierarchical Bayesian model for spatio-temporal areal data, which allows for time varying spatial model-based clustering though product partition models (PPM). We develop efficient MCMC algorithms based on numerical linear algebra, which exploit the sparse structure of the precision matrix of a Gaussian Random Markov Field (GMRF). Finally, we present an application to mobile data, with the goal to model, predict and spatially cluster population density dynamics.

EO308 Room MAL G15 NOVEL APPLICATIONS IN BAYESIAN NONPARAMETRICS Chair: Michele Guindani

E0757: **Exploiting conjugacy to build time dependent feature allocation models**
Presenter: **Raffaele Argiento**, University of Torino, Italy

A flexible approach to build stationary time-dependent processes exploits the concept of conjugacy in a Bayesian framework: the transition law of the process is defined as the predictive distribution of an underlying Bayesian model. If the model is conjugate, the transition kernel can be analytically derived, making the approach particularly appealing. We aim at achieving such a convenient mathematical tractability in the context of completely random measures (CRMs), i.e. when the variables exhibiting a time dependence are CRMs. In order to take advantage of the conjugacy, we consider the wide family of exponential completely random measures. This leads to a simple description of the process which has a autoregressive structure. The proposed process can be straightforwardly employed to extend CRM-based Bayesian nonparametric models such as feature allocation models to time-dependent data. These processes can be applied to problems from modern real life applications in very different fields, from computer science to biology. We develop a dependent latent feature model for the identification of features in images and a dynamic Poisson factor analysis for topic modelling.

E1327: **Marginal Bayesian semiparametric modelling of mismeasured multivariate interval-censored data**
Presenter: **Alejandro Jara**, Pontificia Universidad Catolica de Chile, Chile

Motivated by data gathered in an oral health study, we propose a Bayesian nonparametric approach for population-average modeling of correlated time-to-event data, when the responses can only be determined to lie in an interval obtained from a sequence of examination times and the determination of the occurrence of the event is subject to misclassification. The joint model for the true, unobserved time-to-event data is defined semiparametrically; proportional hazards, proportional odds, and accelerated failure time (proportional quantiles) are all fit and compared. The baseline distribution is modeled as a flexible tail free prior. The joint model is complete by considering a parametric copula function. A general misclassification model is discussed in detail, considering the possibility that different examiners were involved in the assessment of the occurrence

of the events for a given subject across time. We provide empirical evidence that the model can be used to estimate the underlying time-to-event distribution and the misclassification parameters without any external information about the latter parameters.

E1340: **Beta-binomial stick breaking nonparametric prior**
Presenter: **Ramses Mena**, UNAM, Mexico

A new class of stick-breaking nonparametric priors, termed Beta-Binomial process, is proposed. By allowing the underlying stick-breaking sequences to be dependent accordingly to a Beta-Binomial model, an appealing discrete random probability measure arises. Indeed, the resulting class contains the Dirichlet process and the Geometric process priors as particular cases. Tuning the chain's dependence parameter, controls the model's label switching adaptation for a given dataset and a given set of initial parameters. Some properties of the model are discussed and a density estimation algorithm proposed and tested with simulated datasets.

E1409: **Modelling heterogeneous data in Bayesian nonparametrics**
Presenter: **Fabrizio Leisen**, University of Kent, United Kingdom

Some recent advances about dependent Bayesian nonparametric priors will be illustrated.

EO582 Room MAL G16 RECENT ADVANCES IN SEQUENTIAL MONTE CARLO Chair: Jere Koskela

E0401: **Unbiased and consistent nested sampling via sequential Monte Carlo**
Presenter: **Leah South**, Lancaster University, United Kingdom
Co-authors: Robert Salomone, Christopher Drovandi, Dirk Kroese

A new class of sequential Monte Carlo methods is introduced which is called Nested Sampling via Sequential Monte Carlo (NS-SMC), and which reframes the Nested Sampling method of Skilling in terms of sequential Monte Carlo techniques. This new framework allows one to obtain provably consistent estimates of the marginal likelihood and posterior inferences when Markov chain Monte Carlo (MCMC) is used to produce new samples. An additional benefit is that marginal likelihood estimates are also unbiased. For applications of NS–SMC, we give advice on tuning MCMC kernels in an automated manner via a preliminary pilot run, and present a new method for appropriately choosing the number of MCMC repeats at each iteration. A numerical study is conducted where the performance of NS–SMC and temperature–annealed SMC is compared on several challenging and realistic problems.

E0926: **Asymptotic genealogies of SMC methods**
Presenter: **Jere Koskela**, University of Warwick, United Kingdom
Co-authors: Paul Jenkins, Adam Johansen, Dario Spano

It is well known that the genealogy embedded into an SMC algorithm by resampling plays a central role in important questions, such as estimation of variances of SMC estimators, and mixing of conditional SMC schemes. Nevertheless, results on the distribution of genealogies have only been available in the toy setting of particle filters with constant importance weights, in which case genealogies of finite samples of particles converge to the Kingman coalescent in the large ensemble size limit. We will show that the same convergence holds under verifiable assumptions which are typically satisfied by real SMC algorithms on compact state spaces, present a connection between genealogies and effective sample size, and discuss implications for SMC storage cost and variance estimation.

E1111: **Limit theorems for sequential MCMC methods**
Presenter: **Axel Finke**, National University of Singapore, Singapore
Co-authors: Adam Johansen, Arnaud Doucet

Both *sequential Monte Carlo (SMC)* methods ("particle filters") as well as *sequential MCMC* methods constitute classes of algorithms which can be used to approximate expectations with respect to (a sequence of) probability distributions and their normalising constants. While SMC methods sample particles conditionally independently at each time step, sequential MCMC methods sample particles according to an MCMC kernel. The latter have attracted renewed interest recently as they empirically outperform SMC methods in some applications. We establish a strong law of large numbers and a central limit theorem for sequential MCMC methods. In the context of state-space models, we also provide conditions under which sequential MCMC methods can indeed outperform standard SMC methods in terms of asymptotic variance of the corresponding Monte Carlo estimators.

E1138: **Recent progress in genealogies of sequential Monte Carlo algorithms**
Presenter: **Suzie Brown**, University of Warwick, United Kingdom
Co-authors: Adam Johansen, Jere Koskela, Paul Jenkins

The resampling step of SMC induces a particle genealogy: that is, a tree encoding the ancestors in each generation of each of the particles. Information about the structure of the genealogy is of interest in smoothing problems, both for estimating storage costs, and for analysing the severity of ancestral degeneracy. Such genealogies are characterised via an asymptotic analysis as the number of particles tends to infinity. An extension of an existing result to the case of conditional SMC is presented, as well as some work treating more exotic resampling methods including residual resampling and a general class of stochastic rounding-based schemes.

EO082 Room CLO 101 BAYESIAN AND FREQUENTIST APPROACHES WITH BIG DATA Chair: HaiYing Wang

E0733: **High-dimensional interaction detection with false sign rate control**
Presenter: **Daoji Li**, California State University Fullerton, United States

Identifying interaction effects is fundamentally important in many scientific discoveries and contemporary applications, but it is challenging since the number of pairwise interactions increases quadratically with the number of covariates and that of higher-order interactions grows even faster. Although there is a growing literature on interaction detection, little work has been done on the prediction and false sign rate on interaction detection in ultrahigh dimensional regression models. Such a gap is filled. More specifically, we establish some theoretical results on interaction selection for ultrahigh-dimensional quadratic regression models under random design. We prove that the examined method enjoys the same oracle inequalities as the lasso estimator and further admits an explicit bound on the false sign rate. Moreover, the false sign rate can be asymptotically vanishing. These new theoretical characterizations are confirmed by a simulation study.

E0940: **An online updating approach for testing the proportional hazards assumption with streams of survival data**
Presenter: **Yishu Xue**, University of Connecticut, United States
Co-authors: HaiYing Wang, Jun Yan, Elizabeth Schifano

The Cox model, which remains as the first choice in analyzing time-to-event data even for large datasets, relies on the proportional hazards (PH) assumption. When survival data arrive sequentially in chunks, a fast and minimally storage intensive approach to test the PH assumption is desirable. We propose an online updating approach that updates the standard test statistic as each new block of data becomes available, and greatly lightens the computational burden. Under the null hypothesis of PH, the proposed statistic is shown to have the same asymptotic distribution as the standard version computed on the entire data stream with the data blocks pooled into one dataset. In simulation studies, the test and its variant based on most recent data blocks maintain their sizes when the PH assumption holds and have substantial power to detect different violations of the PH assumption. We also show in simulations that our approach can be used successfully with "big data" that exceed a single computer's computational resources. The approach is illustrated with the survival analysis of patients with lymphoma cancer from the Surveillance, Epidemiology, and End

Results Program. The proposed test promptly identified deviation from the PH assumption that was not captured by the test based on the entire data.

E1947: **Bayesian sharp minimax contraction**
Presenter: **Qifan Song**, Purdue University, United States

Shrinkage prior becomes more and more popular in Bayesian modeling for high dimensional sparse problems due to its computational efficiency. Recent works show that a polynomially decaying prior leads to satisfactory posterior asymptotics under regression models. In the literature, statisticians have investigated how the global shrinkage parameter, i.e., the scale parameter, in a heavy tail prior affects the posterior contraction. We explore how the shape of the prior, or more specifically, the polynomial order of the prior tail affects the posterior. We discover that, under sparse normal means models, the polynomial order does affect the multiplicative constant of the posterior contraction rate. More importantly, if the polynomial order is sufficiently close to 1, it will induce the optimal Bayesian posterior convergence, in the sense that the Bayesian contraction rate is sharply minimax, i.e., not only the convergence order, but also the multiplicative constant of the posterior contraction rate are optimal. The above Bayesian sharp minimaxity holds when the global shrinkage parameter follows a deterministic choice which depends on the unknown sparsity s. Therefore, a Beta-prior modeling is further proposed, such that our sharply minimax Bayesian procedure is adaptive to unknown s. Our theoretical discoveries are justified by simulation studies.

EO364 Room CLO 102 RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICS AND RANDOM MATRIX THEORY Chair: Miles Lopes

E1971: **Fast and fair simultaneous confidence bands for functional parameters**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Matthew Reimherr

Quantifying uncertainty using confidence regions is a central goal of statistical inference. Despite this, methodologies for confidence bands in Functional Data Analysis are underdeveloped compared to estimation and hypothesis testing. A major leap forward in this area is made by presenting a new methodology for constructing simultaneous confidence bands for functional parameter estimates. These bands possess a number of striking qualities: (1) they have a nearly closed-form expression, (2) they give nearly exact coverage, (3) they have a finite sample correction, (4) they do not require an estimate of the full covariance of the parameter estimate, and (5) they can be constructed adaptively according to a desired criterion. One option for choosing bands we find especially interesting is the concept of fair bands, where breaches in coverage are equally likely to occur on any two subintervals of the same length, which could be especially useful in longitudinal studies over long time scales. These bands are constructed by integrating and extending tools from Random Field Theory, an area that has yet to overlap with Functional Data Analysis.

E0254: **Time series source separation using dynamic mode decomposition**
Presenter: **Raj Rao Nadakuditi**, University of Michigan, United States
Co-authors: Arvind Prasadan

The dynamic mode decomposition (DMD) extracted dynamic modes are the non-orthogonal eigenvectors of the matrix that best approximates the one-step temporal evolution of the multivariate samples. In the context of dynamic system analysis, the extracted dynamic modes are a generalization of global stability modes. We apply DMD to a data matrix whose rows are linearly independent, additive mixtures of latent time series. We show that when the latent time series are uncorrelated at a lag of one time-step then, in the large sample limit, the recovered dynamic modes will approximate, up to a column-wise normalization, the columns of the mixing matrix. Thus, DMD is a time series blind source separation algorithm in disguise, but is different from closely related second order algorithms such as SOBI and AMUSE. All can unmix mixed ergodic Gaussian time series in a way that ICA fundamentally cannot. We use our insights on single lag DMD to develop a higher-lag extension, analyze the finite sample performance with and without randomly missing data, and identify settings where the higher lag variant can outperform the conventional single lag variant. We validate our results with numerical simulations, and highlight how DMD can be used in change point detection.

E1976: **Asymptotic theory of eigenvectors for large random matrices**
Presenter: **Jinchi Lv**, University of Southern California, United States

Characterizing the exact asymptotic distributions of high-dimensional eigenvectors for large structured random matrices poses important challenges yet can provide useful insights into a range of applications. To this end, we introduce a general framework of asymptotic theory of eigenvectors (ATE) for large structured symmetric random matrices with heterogeneous variances, and establish the asymptotic properties of the spiked eigenvectors and eigenvalues for the scenario of the generalized Wigner matrix noise, where the mean matrix is assumed to have the low-rank structure. Under some mild regularity conditions, we provide the asymptotic expansions for the spiked eigenvalues and show that they are asymptotically normal after some normalization. For the spiked eigenvectors, we establish novel asymptotic expansions for the general linear combination and further show that it is asymptotically normal after some normalization, where the weight vector can be arbitrary. We also provide a more general asymptotic theory for the spiked eigenvectors using the bilinear form. Simulation studies verify the validity of our new theoretical results.

E1978: **High dimensional hypothesis testing via spectral shrinkage**
Presenter: **Haoran Li**, Columbia University, United States
Co-authors: Alexander Aue, Debashis Paul

Inference on high-dimensional data has remained a central topic of statistical research for over a decade. One of the fundamental inferential problems is to test a linear hypothesis under linear models. Under high-dimensional regimes, mainly due to the inconsistency of classical matrix estimators, such as the sample covariance matrix, traditional inferential procedures, such as the likelihood ratio tests (LRT), perform poorly. To correct the inconsistency of LRT, we propose a flexible spectral shrinkage scheme applied to the sample covariance matrix. The spectral shrinkage adjusts the singularity or near-singularity of the sample covariance matrix and meanwhile maintains its eigen-structure. The scheme is shown to compare favorably against a host of existing methods designed to tackle high-dimensional testing problems in a wide range of settings.

EO228 Room Court SEMI AND NON-PARAMETRIC MIXTURE MODELLING Chair: Frans Kanfer

E1033: **Joint modeling of survival and longitudinal quality of life data in palliative care studies**
Presenter: **Zhigang Li**, Department of Biostatistics, University of Florida, United States

Palliative medicine is an interdisciplinary speciality focusing on improving quality of life (QOL) for patients with serious illness and their families. Palliative care programs are widely available or under development at US hospitals. In palliative care studies, often longitudinal QOL and survival data are highly correlated which, in the face of censoring, makes it challenging to properly analyze and interpret terminal QOL trend. Dropout in the study add another level of complication of the problem. To address these issues, we propose a novel statistical approach to jointly model the terminal trend of QOL and survival data accounting for informative dropout. We assess the model through simulation and application to establish a novel modeling approach that could be applied in future palliative care treatment research trials.

E1243: **Mixture of generalised linear models with unknown link**
Presenter: **Sollie Millard**, University of Pretoria, South Africa
Co-authors: Frans Kanfer, Mohammad Arashi

An extension of mixtures of generalised linear models into a semi-parametric mixture setting is considered. The link function is estimated using a non-parametric estimation procedure. This approach allows for more flexibility since the non-parametric link function gives access to a larger

subset of admissible distributions in the exponential family, whilst retaining much of the structure of a generalised linear model. Since the parameter estimates are not directly comparable to the traditional generalised linear model estimates, we consider using ratios of parameters for interpretation. A simulation study is used to evaluate the performance of the estimation procedure. A semi-parametric mixture of binary outcome regression models, based on industry data, is also considered.

E1246: **Generalised partially linear models with unknown link**
Presenter: **Frans Kanfer**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Mohammad Arashi
An extension of the generalised partial linear model (GPLM) is considered. We develop a methodology to estimate regression coefficients in a mixture of GPLMs by approximating the nonparametric component using a B-spline function. We assume no information about the link function, except an initial guess. The link function is updated through the estimation procedure as the model parameters are estimated. A semiparametric setup of the GPLMs and mixture of GPLMs is considered. Performance of the proposed flexible model estimation strategy is validated via an extensive simulation study with an application on industry data.

E1617: **A statistical distribution for simultaneously modeling skewness, kurtosis and bimodality for mixture modelling**
Presenter: **Din Chen**, University of North Carolina, United States
A family of distributions from the cusp catastrophe theory is revitalized. The family was developed in the early 1970s as part of the catastrophe theory in topographic research, which included 7 elementary catastrophes (e.g., fold, cusp, swallowtail, elliptic umbilic, hyperbolic umbilic, butterfly, and parabolic umbilic). These distributions also belong to the classical exponential family, which can be used to statistically analyze data with skewness, kurtosis and bimodal simultaneously for semi and non-parametric mixture modelling. We will show the properties of these distributions and the parameter estimation with the theory of maximum likelihood estimation. We further demonstrate the applications to analyze real data.

EO839 Room Jessel CHANGE-POINTS/ANOMALY DETECTION Chair: Alain Celisse

E1412: **Nonparametric multiple change-point estimation for analyzing large Hi-C data matrices**
Presenter: **Laure Sansonnet**, AgroParisTech / INRA, France
A novel nonparametric approach is proposed for estimating the location of block boundaries (change-points) of non-overlapping blocks in a random symmetric matrix which consists of random variables whose distribution changes from block to block. Our change-point location estimators are based on nonparametric homogeneity tests for matrices. We first provide some theoretical results for these tests. Then, we prove the consistency of our change-point location estimators. Some numerical experiments are also provided in order to support our claims. Finally, our approach is applied to Hi-C data which are used in molecular biology to study the influence of chromosomal conformation on cell function.

E1703: **A statistical method to detect abrupt changes in trees**
Presenter: **Guillem Rigaill**, INRA, France
Co-authors: Solene Thepaut, Nicolas Verzelen
The problem of detecting multiple changes in the mean of the nodes of a tree is studied. This problem is motivated by an application in ecology, where diversity measurements are made at n points of a river system. The river system is represented as a tree. The goal is to detect sub-trees where the diversity is abnormally high or low. We propose to infer the signal and the position of the changes by minimizing a penalized empirical risk. We propose a penalty satisfying a non-asymptotic oracle inequality. We propose a pruned dynamic programming algorithms to solve this problem. We empirically show that their complexity is on average $O(n^2)$ or less with n the number of nodes of the tree. We tested our approach on simulations and used it on our ecological dataset.

E1806: **Detecting abrupt changes in correlated time-series**
Presenter: **Gaetano Romano**, Lancaster University, United Kingdom
Co-authors: Guillem Rigaill, Paul Fearnhead, Vincent Runge
Change-Point analysis has been of major interest in recent times. Current algorithms for detecting changes in mean often struggle in the presence of auto-correlated noise, or in situations where the mean can vary locally between abrupt changes of interest. Default implementations of the algorithms in these scenarios will lead to detecting many spurious changes. This can be corrected for, but with a resulting loss of power. We develop principled statistical approaches to estimate changes under both these scenarios. These are based on maximising a penalised likelihood for appropriate models for the data. Estimating the change-points locations is non-trivial as it involves a solving a complicated, non-convex, optimisation problem. We show how to extend recent dynamic programming ideas to obtain exact solutions of this optimisation in (empirically) close to linear time in the number of observations. Our method is shown to out-perform alternative methods both on simulated and real-data.

E1853: **Efficient change-in-slope optimal partitioning algorithm in a finite-size parameter space**
Presenter: **Vincent Runge**, Evry Paris-Saclay University, France
Co-authors: Nicolas Deschamps de Boishebert, Marco Pascucci, Guillem Rigaill
The problem of detecting change-points in univariate time series is considered by fitting a piecewise linear continuous signal. Values for beginning and ending points of each segment are restricted to a finite set of size m. Using this finite parameter space, we write a dynamic programming algorithm with time complexity $O(m^2n^2)$ (for n data points) which is similar to the optimal partitioning approach. Some pruning strategies can be added to reduce the constant before n^2. As for functional pruning optimal partitioning, we are able to constrain the inference to an increasing signal (isotonic constraint) to a unimodal signal (up-down constraint) or to a robust (to outliers) signal. Our algorithm was initially developed to analyse beat-per-minute time series in order to build a simplified continuous signal with integer change-points which is the right level of information for musicians.

EO614 Room MAL 152 METHODS FOR UNDERSTANDING NETWORK DATA STRUCTURES Chair: Nynke Niezink

E0453: **On spectral embedding performance and elucidating network structure in stochastic blockmodel graphs**
Presenter: **Joshua Cape**, University of Michigan, United States
Statistical inference on graphs often proceeds via spectral methods involving low-dimensional embeddings of matrix-valued graph representations such as the graph Laplacian or adjacency matrix. We analyze the asymptotic information-theoretic relative performance of Laplacian spectral embedding and adjacency spectral embedding for block assignment recovery in stochastic blockmodel graphs by way of Chernoff information. We investigate the relationship between spectral embedding performance and underlying network structure (e.g., homogeneity, affinity, core-periphery, (un)balancedness) via a comprehensive treatment of the two-block stochastic blockmodel and the class of K-blockmodels exhibiting homogeneous balanced affinity structure. Our findings support the claim that, for a particular notion of sparsity, loosely speaking, "Laplacian spectral embedding favors relatively sparse graphs, whereas adjacency spectral embedding favors not-too-sparse graphs." We also provide evidence in support of the claim that "adjacency spectral embedding favors core-periphery network structure".

E0972: **Change point detection for networks with dynamic community structure**
Presenter: **David Choi**, Carnegie Mellon University, United States

A method is proposed for change point detection in networks that display time-varying community structure, and a bootstrap-based confidence interval to characterize the change in community membership at each change point. To find multiple change points, a simple extension to existing search methods is also proposed, which combines aspects of greedy methods (such as wild binary segmentation) and global optimization by dynamic programming.

E1017: **Consistent polynomial-time unseeded graph matching for Lipschitz graphons**
Presenter: **Yuan Zhang**, Ohio State University, United States

A consistent polynomial-time method is proposed for the unseeded node matching problem for networks with smooth underlying structures. Despite widely conjectured by the research community that the structured graph matching problem to be significantly easier than its worst case counterpart, well-known to be NP-hard, the statistical version of the problem has stood a challenge that resisted any solution both provable and polynomial-time. The closest existing work requires quasi-polynomial time. Our method is based on the latest advances in graphon estimation techniques and analysis on the concentration of empirical Wasserstein distances. Its core is a simple yet unconventional sampling-and-matching scheme that reduces the problem from unseeded to seeded. Our method allows flexible efficiencies, is convenient to analyze and potentially can be extended to more general settings. Our work enables a rich variety of subsequent estimations and inferences.

E0269: **Change point detection in partial correlation networks**
Presenter: **Jessica Wai Yin Leung**, University of Sydney, Australia
Co-authors: Dmytro Matsypura

The focus is on the problem of change point detection in the structure of similarity network over time. The correlation matrix is commonly used in constructing similarity networks. Yet, it is well known that it suffers from the illusion of spurious relationships within the system. Therefore, we adopt the partial correlation matrix as an alternative similarity measure and develop a statistical testing procedure that is tailored to such construct with minimal distributional assumptions. Simulation results show that our test has high statistical power across different sample sizes and dimensions.

EO502 Room MAL 153 NEW METHODS FOR COMPLEX DATA ANALYSIS Chair: Zhihua Su

E0713: **Latent simplex position model: Multiview clustering of high dimensional data**
Presenter: **Leo Duan**, University of Florida, United States

High dimensional data often contain multiple facets, and several clustering patterns (views) can co-exist under different feature subspaces. While multi-view clustering algorithms were proposed, the uncertainty quantification remains difficult — a particular challenge is in the high complexity of estimating the cluster assignment probability under each view, or/and to efficiently share information across views. We propose an empirical Bayes approach — viewing the similarity matrices generated over subspaces as rough first-stage estimates for co-assignment probabilities, in its Kullback-Leibler neighborhood we obtain a refined low-rank soft cluster graph, formed by the pairwise product of simplex coordinates. Interestingly, each simplex coordinate directly encodes the cluster assignment uncertainty. For multi-view clustering, we equip each similarity matrix with a mixed membership over a small number of latent views, leading to effective dimension reduction. With a high model flexibility, the estimation can be succinctly re-parameterized as a continuous optimization problem, hence enjoys gradient-based computation. Theory establishes the connection of this model to random cluster graph under multiple views. Compared to single-view clustering approaches, substantially more interpretable results are obtained when clustering brains from human traumatic brain injury study, using high-dimensional gene expression data.

E1332: **Element-wise estimation error of a total variation regularized estimator for change point detection**
Presenter: **Teng Zhang**, University of Central Florida, United States

The total variation regularized 2 estimator (fused lasso) is studied in the setting of a change point detection problem. Compared with existing works that focus on the sum of squared estimation errors, we give bound on the element-wise estimation error. The bound is nearly optimal in the sense that the sum of squared error matches the best existing result, up to a logarithmic factor. This analysis of the element-wise estimation error allows a screening method that can approximately detect all the change points. We also generalize this method to the muitivariate setting, i.e., to the problem of group fused lasso.

E0847: **Envelope quantile regression**
Presenter: **Guangyu Zhu**, University of Rhode Island, United States
Co-authors: Lan Wang, Zhihua Su, Shanshan Ding

Quantile regression offers a valuable complement of classical mean regression for robust and comprehensive data analysis in a variety of applications. We propose a novel envelope quantile regression method (EQR) that adapts a nascent technique called enveloping to improve the efficiency of standard quantile regression. The new method aims to identify material and immaterial information in a quantile regression model and use only the material information for estimation. By excluding the immaterial part, the EQR method has the potential to substantially reduce the estimation variability with standard quantile regression. Unlike existing envelop model approaches which mainly rely on the likelihood framework, our proposed estimator is defined through a set of nonsmooth estimating equations. We facilitate the estimation via the generalized method of moments and derive the asymptotic normality of the proposed estimator by applying empirical process techniques. Furthermore, we establish that EQR is asymptotically more efficient than (or at least as asymptotically efficient as) the standard quantile regression estimators without imposing stringent conditions. Hence, we advance the envelope model theory to general distribution-free settings. We demonstrate the effectiveness of the proposed method via Monte-Carlo simulations and a real data example.

E1394: **A Bayesian approach to envelope quantile regression**
Presenter: **Zhihua Su**, University of Florida, United States

The envelope model is a nascent construct that aims to increase efficiency in multivariate analysis. It has been used in many contexts including linear regression, generalized linear models, matrix or tensor variate regression, reduced rank regression, and quantile regression, and has showed the potential to provide substantial efficiency gains. Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is sparse. The objective is to propose a Bayesian approach for envelope quantile regression. The proposed approach has straightforward interpretation of model parameters and allows easy incorporation of prior information. We provide a simple block Metropolis-within-Gibbs MCMC sampler for practical implementation of our method. Simulations and data examples are included for illustration.

EO506 Room MAL 254 METHODS AND COMPUTATION FOR MODELING DATA IN SPACE AND TIME Chair: Francois Bachoc

E0536: **Random domain decomposition for spatial prediction of manifold-valued data**
Presenter: **Davide Pigoli**, Kingś College London, United Kingdom
Co-authors: Alessandra Menafoglio, Piercesare Secchi

Data taking value on a Riemannian manifold and observed over a complex spatial domain are becoming more and more frequent in applications, e.g. in environmental sciences and in geoscience. The analysis of these data needs to rely on local models to account for the non stationarity of the generating random process, the non linearity of the manifold and the complex topology of the domain. A method is illustrated to predict a spatial field taking value on a smooth Riemannian manifold. A Random Domain Decomposition of the spatial domain will be used to estimate

an ensemble of local models which will be eventually aggregated through Fréchet averaging. The proposed algorithm will be first described in complete generality and then applied to the case of positive definite matrices. As an illustrative case study, we will report the analysis of covariance matrices for an environmental dataset observed over the Chesapeake Bay (USA).

E0792: **Spatio-temporal modelling in environmental and ecological systems**
Presenter: **Claire Miller**, University of Glasgow, United Kingdom

Multiple potential data sources exist to aid in the monitoring and management of environmental and ecological systems. Data include, for example, those collected from long-term monitoring programmes, citizen science data, automatic sensor monitoring data and data from processed satellite retrievals. However, information gaps still exist making global challenges such as water pollution difficult to address. For example, the 2018 UN Water SDG 6 Synthesis Report suggests that the global data currently collected through the SDG process do not reflect the general state or trends known about freshwater ecosystems from other data sources. Where data do exist, the challenge lies in appropriately combining the available data streams to fill the knowledge gaps by providing improved estimation and prediction of, for example, water quality. The purpose is to give examples of the statistical methodological and computational challenges presented in such a context including combining data streams which are of different spatial and temporal support, identifying and accounting for potential bias and uncertainty in data retrievals, reducing dimensionality (while accounting for sparse data) and accounting for highly correlated nested spatial scales. Methodological developments and examples from recent work, including the GloboLakes and Hydroscape projects, will be presented.

E1147: **Non-parametric multivariate spatio-temporal model for global climate ensembles**
Presenter: **Matthew Edwards**, Newcastle University, United Kingdom

The variables of global climate ensembles have complex spatial and multivariate dependencies that vary over the globe. It is important to capture these dependencies in a stochastic model for accurate simulation and prediction. However, specifying and fitting such a model is incredibly difficult. Marginally parameterised models have been used to overcome the difficulty of fitting as they allow subsets of the parameters to be estimated with subsets of the data very efficiently. However, previous specifications have struggled to capture variations in spatial and multivariate dependencies over the globe. We propose a non-parameteric version of these models that uses a novel non-stationary periodic AR process over lines of longitude in combination with a non-stationary AR process over lines of latitude to capture these dependencies.

E1496: **Challenges and solutions for multiscale global temperature reconstruction**
Presenter: **Finn Lindgren**, University of Edinburgh, United Kingdom

As part of the EUSTACE project, an ambitious aim of constructing a historical global and daily temperature reconstruction system using recent advances in statistical modelling and computational methodology was envisioned. To approach this ambitious mission, models needed to be developed and investigated, to design a system that could integrate diverse observation systems, station biases, non-Gaussian temperature distributions, and diverse spatial and temporal correlation scales. At the same time, these high-dimensional models required novel computational approaches. We will present some of the challenges and solutions that were developed, with emphasis on the major modelling and computational challenges in taking theoretical models and scaling the associated computational methods from one hundred thousand dimensions up by six orders of magnitude in size, leveraging parallel computing and sparse iterative numerical solvers.

EO256 Room Senate MODERN TOPICS IN STATISTICS OF EXTREMES Chair: Yuri Goegebeur

E0365: **Robust nonparametric estimation of the conditionaltail dependence coefficient**
Presenter: **Armelle Guillou**, Strasbourg, France
Co-authors: Yuri Goegebeur, Nguyen Khanh Le Ho, Jing Qin

Robust and nonparametric estimation of the coefficient of tail dependence is considered in presence of random covariates. The estimator is obtained by fitting the extended Pareto distribution locally to properly transformed bivariate observations using the minimum density power divergence criterion. We establish convergence in probability and asymptotic normality of the proposed estimator under some regularity conditions. The finite sample performance is evaluated with a small simulation experiment, and the practical applicability of the method is illustrated on a real dataset of air pollution measurements.

E0485: **What hides behind an extreme currency demand: Bayesian conditionally heteroscedastic extremes**
Presenter: **Miguel de Carvalho**, University of Edinburgh, United Kingdom
Co-authors: Junho Lee, Antonio Rua

Bayesian inference methods will be introduced for conditionally heteroscedastic extremes. The proposed model is based on an tail index regression and on a proportional tails model, and can be used for assessing how the magnitude and frequency of the extreme values can change along with a covariate. We start with the unconditional setting for estimating the tail index and the scedasis function and show that the proposed inference methods for the scedasis density—based on a Bernstein–Dirichlet prior—perform well in Monte Carlo simulation studies, are exact apart from Monte Carlo error, and have full support on the space of all continuous scedasis functions. We then extend the proposed methods to the conditional setting using dependent Bernstein-Dirichlet process. We resort to the proposed methodologies to examine an extreme currency demand in Portugal. The signatures of the fitted scedasis densities of extreme currency demand—over different denominations—reveal some interesting insights on the dynamics governing currency demand over periods of economic stress.

E0487: **On a relationship between randomly and non-randomly thresholded empirical average excesses for heavy tails**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France

Motivated by theoretical similarities between the classical Hill estimator of the tail index of a heavy-tailed distribution and one of its pseudo-estimator versions featuring a non-random threshold, we show a novel asymptotic representation of a class of empirical average excesses above a high random threshold, expressed in terms of order statistics, using their counterparts based on a suitable non-random threshold, which are sums of independent and identically distributed random variables. As a consequence, the analysis of the joint convergence of such empirical average excesses essentially boils down to a combination of Lyapunov's central limit theorem and the Cramer-Wold device. We illustrate how this allows us to improve upon, as well as produce conceptually simpler proofs of, very recent results about the joint convergence of marginal Hill estimators for a random vector with heavy-tailed marginal distributions. These results are then applied to the proof of a convergence result for a tail index estimator when the heavy-tailed variable of interest is randomly right-truncated.

E1582: **Conditional marginal expected shortfall**
Presenter: **Nguyen Khanh Le Ho**, University of Southern Denmark, Denmark
Co-authors: Yuri Goegebeur, Armelle Guillou , Jing Qin

In the context of bivariate random variables $(Y^{(1)}, Y^{(2)})$, the marginal expected shortfall, defined as $\mathbb{E}(Y^{(1)}|Y^{(2)} \geq Q_2(1-p))$ for p small, where Q_2 denotes the quantile function of $Y^{(2)}$, is an important risk measure, which finds applications in areas like, e.g., finance and environmental science. We consider estimation of the marginal expected shortfall when the random variables of main interest $(Y^{(1)}, Y^{(2)})$ are observed together with a random covariate X, leading to the concept of the conditional marginal expected shortfall. The asymptotic behavior of an estimator for this conditional marginal expected shortfall is studied for a wide class of bivariate distributions, with heavy-tailed marginal distributions, and where p tends to zero at an intermediate rate. The finite sample performance is evaluated on a small simulation experiment.

EO504 Room CLO 203 RECENT DEVELOPMENTS IN OPTIMAL EXPERIMENTAL DESIGNS Chair: Subir Ghosh

E0719: **Optimal design, lagrangian and linear models theories: A fusion**
Presenter: **Ben Torsney**, University of Glasgow, United Kingdom

The purpose is optimizing criterion $f(p)$ (p nonneg vector) subject to several equality constraints: $Ap = b$. (wlog b is nonneg.) Lagrangian Theory requires that, at optima, partial derivatives be exactly linear in Lagrange Multipliers (LMs). So partial derivatives, viewed as response variables, must exactly satisfy a Linear Model with LMs as parameters. This is a model without errors, implying zero residuals. Residuals appear to play the role of directional derivatives, as defined for optimal designs when $A = (1,1,..,1)$, $b = 1$. Further we extend a class of multiplicative algorithms, designed for the latter case, to our problem. The algorithm has two steps: (i) a multiplicative one, multiplying the current values of the components of p by an increasing function of partial or directional derivatives; (ii) a scaling step under which the products formed in (i) are scaled to meet the summation to one constraint. Step (i) readily extends to our problem, while the more challenging step (ii) has been surmounted in some examples, but needs further development. Results in two main areas will be reported: (a) constraints on multinomial models, given data from multidimensional contingency tables, defined by fixed marginal distributions or, for square tables, hypotheses of marginal homogeneity; (b) optimal approximate designs subject to cost constraints, or, subject to given marginal approximate designs; in these cases an optimal p can have zero components.

E1210: **Response adaptive designs with asymptotic optimality**
Presenter: **Yanqing Yi**, Memorial University of Newfoundland, Canada

The asymptotic optimality of statistical inference for response adaptive designs, which have the ethical advantages over the traditional methods for clinical trials, is discussed. The upper bound of statistical power of asymptotically level tests is derived and the Wald statistic is shown to be asymptotically optimal in terms of achieving the upper bound. The rates of coverage error probability of the confidence interval are proven to depend on the convergence rate of the allocation proportions for non-normally distributed responses. When the response density functions are normal density functions, it is proven that the coverage error probability and the type I error rate has the order of n^{-1}.

E1289: **Causation, bias and optimal experimental design**
Presenter: **Henry Wynn**, London School of Economics, United Kingdom

One of the profound issues in causal modelling, particularly in socio-medical areas is the protection of causal models from bias from different sources, the most problematical being the inability to do controlled experiments, often labelled as the "absence of counterfactuals". Building on previous work, we study optimal experimental designs for models in which either there is a well-defined bias term or in which we guard against vaguer possible sources of bias. We adopt a cooperative game-theoretic approach in which a hypothetical player Alice has "ownership" of the main model, and another player "Bob" tries to eliminate the effect of bias. The standard randomized control methods falls into this category. Motivated by minimax justifications of randomization we are, in certain circumstances, able to establish the existence of Nash equilibrium. This then leads to new optimal design criteria. Complex causal models may be non-linear with additional complexity. Removing or guarding against bias can then be seen as providing a protective Markov "blanket".

E0563: **Search for optimality in the estimation of variance components**
Presenter: **Subir Ghosh**, University of California, United States

The optimal estimators of variance components with closed form expressions are often difficult or impossible to obtain. In such situations, the search for estimators near to the optimal estimators are presented satisfying the condition of unbiasedness. In addition, the estimators have exact closed form expressions. When the closed form expressions are unavailable for the optimum estimators of variance components, the proposed near optimal estimators are obtained as approximation to the optimal estimators. Illustrative examples are presented.

EO134 Room CLO 204 STATISTICS IN NEUROSCIENCE II Chair: Russell Shinohara

E0217: **Inter-modal coupling analysis**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Russell Shinohara

Local cortical coupling is a subject-specific measure of the spatially varying relationship between cortical thickness and sulcal depth. Although it is a promising first step towards understanding local covariance patterns between two image-derived measurements, a more general coupling framework that can accommodate multiple volumetric imaging modalities is warranted. We first introduce Inter-Modal Coupling (IMCo), an analogue of local coupling in volumetric space that can be used to produce subject-level, spatially varying feature maps derived from two volumetric imaging modalities. We then leverage IMCo to address partial volume effects when studying localized relationships between gray matter density and cerebral blood flow (CBF) among participants in the Philadelphia Neurodevelopmental Cohort. We also develop a generalized estimating equation approach to study spatial variation in multi-modal image relationships at the population level.

E0247: **Leveraging spatial dependencies on the cortical surface to improve estimation of subject-level brain organization**
Presenter: **Amanda Mejia**, Indiana University, United States

A primary objective in many functional magnetic resonance imaging (fMRI) studies is localization of functional areas, regions of the brain exhibiting synchronous activity. This is true of both task and resting-state studies, where a goal is to identify regions that coactivate in the absence of a specific task. fMRI data is composed of small voxels, whose contributions to a functional area form a spatial field. These fields exhibit strong spatial dependence, since neighboring voxels tend to exhibit similar behavior. However, models used to estimate these functional areas have often considered voxels to be independent, resulting in loss of efficiency and power. Spatial Bayesian models have been proposed as a way to account for spatial dependence, but the complex dependence structure of fMRI is difficult to accurately represent with models simple enough to be tractable in high dimensions. A promising alternative is to use cortical surface fMRI (cs-fMRI), which projects the cortical gray matter to a triangular mesh manifold surface, where the spatial dependence structure is simplified. We propose leveraging spatial dependencies along the cortical surface through a Bayesian modeling framework with stochastic partial differential equation (SPDE) spatial process priors, which are built on a triangular mesh. We demonstrate this approach through task and resting-state fMRI studies and quantify its benefits through reliability studies using data from the Human Connectome Project.

E0656: **What the proportional recovery rule is (and is not): Methodological and statistical considerations**
Presenter: **Jeff Goldsmith**, Columbia University, United States

In 2008, it was proposed that the magnitude of recovery from non-severe impairment over the first 3-6 months after stroke, as measured with the Fugl-Meyer Assessment (FMA), is approximately 0.7*initial impairment (proportional recovery). In contrast to patients with non-severe hemiparesis, about 30% of patients with severe initial paresis do not show such recovery (non-recoverers). Based on these findings it was suggested that the proportional recovery rule (PRR) was a manifestation of a spontaneous mechanism that is present in all patients with mild-to-moderate paresis but only in some with severe paresis. Since the introduction of the proportional recovery (PR) rule, it has subsequently been applied to other motor outcomes and to non-motor deficits. This more general investigation of the PR rule has led to inconsistencies in its formulation and application, making it difficult to draw conclusions across studies and precipitating some cogent criticism. We conduct a detailed comparison of the different studies reporting proportional recovery and, where appropriate, critique statistical methodology. On balance, we conclude that existing

data in aggregate are largely consistent with the PRR as a population-level mechanism for FMA-UE recovery; recent reports of its demise are exaggerated, as these excessively focus on less conclusive subject-level predictions.

E1571: **Modeling brain connectivity in real time**

Presenter: **Hernando Ombao**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

The motivation comes from the problem of characterizing multi-scale changes in multivariate time series resulting from an external stimulus or shock to the system. One particular goal is to develop a method that can track real-time changes in dependence. A quick overview of the classical measures will be covered: coherence, partial coherence and dual-frequency coherence and then introduce some non-stationary generalizations of these (in particular, the evolutionary dual-frequency coherence). We then discuss partial directed coherence which, unlike the previously mentioned measures, can capture directionality between components under the framework of vector autoregressive processes. Some of the real-time techniques for estimating the different measures of connectivity and for extracting low-dimensional signal summaries will be covered. These methods will be critical for understanding biofeedback and adjusting the stimuli adaptively during the experiment. These methods will be applied to various brain signals to track dynamic changes in connectivity in an experiment that seeks to find associations between brain physiological signals and creative thinking.

EO753 Room MAL 251 POLYNOMIALS IN STATISTICS Chair: Denys Pommeret

E0623: **Polynomials and risk models**

Presenter: **Claude Lefevre**, Universite Libre de Bruxelles, Belgium

The goal is to exhibit the underlying algebraic polynomial structure in the study of stochastic risk models. This is highlighted for two different topics of applied probability: the probability of ruin in insurance theory and the final size distribution in epidemic modeling. The polynomials involved are of a similar but different nature. In insurance, they are of Appell type, while in epidemics, they are of Abel-Gontcharoff type. They can also depend on several variables to take into account effects of heterogeneity. The basic properties of both polynomial families are discussed and then used to analyze the two problems of interest.

E0614: **Projection estimation of multivariate copula densities**

Presenter: **Yves Ismael Ngounou Bakam**, Institute of Mathematics of Marseille, France

Co-authors: Denys Pommeret

A non parametric copula estimator is proposed based on an orthogonal expansion of its density. We investigate the asymptotic properties. From the theoretical point of view, we study the performance of the procedure in the minimax sense and maxiset approach. We provide a convenient way to select the smoothing parameter, the relevant orthogonal basis and reference measure. Numerical analysis and comparisons show the potential of this procedure. An illustration is presented through a real actuarial data set.

E1067: **Testing equality of two processes**

Presenter: **Denys Pommeret**, Institute of Mathematics of Marseille, France

Co-authors: Laurence Reboul, Anne Francoise Yao

The comparison of stationary processes is still an important topic with many applications as financial series, environmental data, and so on. The main difficulty comes from testing all the multivariate joint distributions to accept the equality of processes. By simplifying, many papers are concerned by the comparison of the marginal distributions of two processes, but they do not give a satisfactory answer. We propose to test the equality of the distribution of two stationary processes, possibly paired, with short or long memory, by considering all their multivariate joint distributions. Their d dimensional joint densities are expressed in a multivariate orthogonal basis and their k first coefficients are compared. The dimension d and the number k of coefficients can growth with the sample size and are simultaneously and automatically selected by a two step data driven procedure. A simulation study shows the good behavior of the test procedure and real data sets on financial assets of the US economic sectors are examined.

E1195: **Orthogonal polynomial expansion with applications to insurance**

Presenter: **Pierre-O Goffard**, Universita Lyon 1, France

A numerical method to approximate the probability density function of a nonnegative random variable is considered. It relies on an expansion in terms of the gamma density and its associated orthogonal polynomials. The method comes in very handy when applied to risk measure evaluation. The extension to statistical estimation is discussed as a perspective.

EO512 Room MAL 252 RECENT ADVANCES IN REGRESSION AND CLASSIFICATION FOR HIGH DIMENSIONAL DATA Chair: Ziqi Chen

E0957: **A general framework for sparse learning in reproducing kernel Hilbert space**

Presenter: **Xin He**, Shanghai University of Finance and Economics, China

Co-authors: Junhui Wang

Sparse learning aims to learn the sparse structure of the true target function in various scenarios, which plays a crucial role in high-dimensional data analysis. A general framework is proposed for learning sparsity in M-estimators in a reproducing kernel Hilbert space (RKHS). The M-estimator admits a wide range of loss functions, and thus includes many scenarios as its special cases, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification. The proposed framework is motivated by the properties of RKHS, and its asymptotic estimation and selection consistencies are established without any explicit model specification. Its key advantages are that it works for a general loss function, admits general dependence structure, with theoretical guarantee, and allows for efficient computation. The superior performance of the proposed framework is also supported by a variety of simulated examples and a real application in the human breast cancer study (GSE20194).

E0998: **Surface functional models**

Presenter: **Ziqi Chen**, Central South University, China

The aim is to develop a new framework of surface functional models for surface functional data which contains repeated observations in two domains (typically, time-location). The primary problem of interest is to investigate the relationship between a response and the two domains, where the numbers of observations in both domains within a subject may be diverging. The surface functional models are far beyond the multivariate functional models with two-dimensional predictor variables. Unprecedented complexity presented in the surface functional models, such as possibly distinctive sampling designs and the dependence between the two domains, makes the theoretical investigation of the resulting estimator challenging. We provide a comprehensive investigation of the asymptotic properties of the local linear estimator of the mean function based on a general weighting scheme, including equal weight (EW), direction-to-denseness weight (DDW) and subject-to-denseness weight (SDW), as special cases.

E1120: **Linear discriminant analysis with high dimensional mixed variables**

Presenter: **Zhongqing Yang**, The Hong Kong Polytechnic University, Hong Kong

Co-authors: Binyan Jiang, Cheng Wang, Chenlei Leng

With the rapid development of modern measurement technologies, datasets containing both discrete and continuous variables are more and more commonly seen in different areas. The dimensions of the discrete and continuous variables are often be very high. Though discriminant analysis

for mixed variables under the traditional fixed dimension setting has been well studied since the 80's, promising approaches taking into account both the high dimensionality and the mixing nature of the data sets are still missing. We aim to develop a simple yet useful classification rule that addresses both the high dimensionality and the mixing nature of the variables simultaneously. Our framework is built on a location model, under which we further propose a semiparametric formulation for the optimal Bayes rule. We show that the optimal classification direction and the intercept in the optimal rule can be estimated separately. Efficient direct estimation schemes are then developed to obtain consistent estimators of the discriminant components. Asymptotic results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by simulation and real data studies.

E1674: **Multivariate functional subspace classification for high-dimensional longitudinal data and its application**
Presenter: **Tatsuya Fukuda**, Chuo University, Japan
Co-authors: Toshihiro Misumi, Yoshihiko Maesono, Sadanori Konishi

Classification for high-dimensional longitudinal data with multiple classes plays an important role in diverse fields of the natural and social sciences. The subspace method known as the class-featuring information compression (CLAFIC) based on principal component analysis is a useful tool for classifying and representing patterns, and a number of applications of CLAFIC method have been reported in character recognition, speech recognition, image analysis, etc. However, a disadvantage of this procedure is that it may not be applied to longitudinal studies where subjects are measured at different time points. In order to overcome this issue, we propose a novel classification method for high-dimensional longitudinal data with multiple classes by extending CLAFIC method, and we call it multivariate functional subspace method (mFSM). The mFSM can be used to classify an unlabeled data by measuring the distance between the data and a subspace for each class, obtained by a multivariate functional principal component analysis. Since the accuracy of mFSM based classifier deeply depends on the dimension of subspaces, we consider the problem of selecting the optimal dimension of subspaces. The performance of proposed method is evaluated through a simulation study, and we present the results of the analysis of handwritten number data.

EO292 Room MAL 253 SPATIAL MODELS FOR INFERENCE ON EPIDEMIOLOGICAL AND SOCIAL INDICATORS Chair: Veronica Berrocal

E0659: **Incorporating population-at-risk uncertainty into disease mapping models**
Presenter: **Lance Waller**, Emory University, United States

Small area health studies typically rely on multiple sources of data to define local disease risk with health registries providing information on observed outcomes of interest and census (or other administrative) data defining the numbers of individuals of risk and potential sociodemographic risk factors. In the United States, some demographic aspects (e.g., age, race, and sex) are available for census small areas (tracts, block groups, and blocks) from the U.S. Census Short Form, while some (e.g., economic variables, housing) were historically available for census small areas from the U.S. Census Long Form which was replaced by the American Community Survey (ACS) in 2010. The ACS provides a rolling sample of the U.S. population and provides small area estimates for specific time periods and with measures of error. We provide a brief review of the role of census demographics in small area health studies, define the available data, and illustrate the impact on small area health studies using data from the 2000 and 2010 U.S. Census and small area health statistics from the state of Georgia, with particular attention on incorporating the reported ACS error into typical models of small area health effects.

E0717: **A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics**
Presenter: **Emanuele Giorgi**, Lancaster University, United Kingdom

Multiple diagnostic tests are often used due to limited resources or because they provide complementary information on the epidemiology of a disease under investigation. Existing statistical methods to combine prevalence data from multiple diagnostics ignore the potential over-dispersion induced by the spatial correlations in the data. To address this issue, we develop a geostatistical framework that allows for joint modelling of data from multiple diagnostics by considering two main classes of inferential problems: (1) to predict prevalence for a gold-standard diagnostic using low-cost and potentially biased alternative tests; (2) to carry out joint prediction of prevalence from multiple tests. We apply the proposed framework to two case studies: mapping Loa loa prevalence in Central and West Africa, using miscroscopy and a questionnaire-based test called RAPLOA; mapping Plasmodium falciparum malaria prevalence in the highlands of Western Kenya using polymerase chain reaction and a rapid diagnostic test. We also develop a Monte Carlo procedure based on the variogram in order to identify parsimonious geostatistical models that are compatible with the data. Our study highlights (i) the importance of accounting for diagnostic-specific residual spatial variation and (ii) the benefits accrued from joint geostatistical modelling so as to deliver more reliable and precise inferences on disease prevalence.

E1117: **Bayesian disaggregation of spatio-temporal community indicators estimated via surveys**
Presenter: **Veronica Berrocal**, University of Michigan, United States

The American Community Survey (ACS) is an ongoing survey administered by the US Census Bureau which collects social, economic, and other community data. ACS estimates are released annually, with varying spatial and temporal resolution: 5-year time periods refer to smaller municipal subdivisions, while 1-year time periods refer to larger areas. Although these estimates contain important community information and are often used in social and health studies, their spatial and temporal resolution pose various challenges: the 5-year ACS estimates might be temporally misaligned with finely resolved outcome data, whereas the coarser 1-year estimates are likely spatially misaligned with finely resolved outcome data. We present a Bayesian hierarchical model that leverages both 1-year and 5-year ACS data and accounts for the survey sampling design to obtain estimates of community indicators at any given spatial and temporal resolution. The disaggregation is achieved by introducing a latent, point-referenced process, in turn modeled using a multi-resolution basis function expansion, which is linked to the ACS data via a stochastic model that accounts also for the survey design.

E1492: **Estimating the risk of dowry death in India: Avoiding confounding to evaluate its association with sex ratio**
Presenter: **Maria Dolores Ugarte**, Universidad Publica de Navarra, Spain

Dowry death is a form of crime against women specific to India. It is a complex phenomenon and looking for potential risk factors is a crucial matter However, when using ecological spatio-temporal models to evaluate the association between potential risk factors and the dowry death risk, it is usually difficult to separate the covariate effects from the spatial, temporal, and spatio-temporal random effects. To deal with this issue (confounding) we consider two possibilities: extending the spatial restricted regression to the spatio-temporal setting and exploring the use of constraints. Both proposals will be compared in terms of fit and fixed effects estimates. The main aim is to assess the association between dowry deaths risks and sex ratio in Uttar Pradesh (the most populated Indian state).

EO618 Room SH349 TOPICS IN TIME SERIES ANALYSIS Chair: Konstantinos Fokianos

E0523: **Categorical changepoint detection for activity sequences**
Presenter: **Jessica Gillam**, Lancaster University, United Kingdom
Co-authors: Rebecca Killick, Simon Taylor, Jamie-Leigh Chapman

Age UK state, as of April 2019 there are close to 12 million people in the UK aged 65 or over, of which 3.8 million live by themselves. Around 40% have a long term condition and approximately 30% need help with at least one daily activity. There is an increasing body of research that indicates changes in daily routine signal a change in health and well-being. This project is in collaboration with Howz who are using sensors on common household appliances to automatically detect changes in well-being. The goal is to investigate which appliances are being used and in what order to detect a change in behaviour whilst allowing for an individual's daily variation. For example, a move from using the hob to microwave

might indicate a change in confidence levels in preparing meals. Classic pattern recognition literature finds it difficult to classify changes in noisy categorical data, whilst the majority of the changepoint literature focus on numerical time series. We present a new method to detecting changes in patterns of behaviour using activity event based sequences with the aim to detect changes on a day-to-day basis.

E0646: **Wavelet spectral analysis of non-stationary circadian rhythms**
Presenter: **Marina Knight**, University of York, United Kingdom

Rhythmic data are ubiquitous in the life sciences. Biologists need reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in a rhythmic signal. When these signals display non-stationary behaviour, as is common in many biological systems, the established methodologies may be misleading. Therefore, there is a real need for new methodology that enables the formal comparison of non-stationary processes. As circadian behaviour is best understood in the spectral domain, we develop novel hypothesis testing procedures in the (wavelet) spectral domain, embedding replicate information when available. The data are modelled as realisations of locally stationary wavelet processes, allowing us to define and rigorously estimate their evolutionary wavelet spectra. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. We demonstrate the advantages of our methodology over alternative approaches, by means of a comprehensive simulation study and real data applications, using both published and newly generated circadian datasets.

E0963: **Semiparametric modeling of multiple quantiles**
Presenter: **Alessandra Luati**, University of Bologna, Italy
Co-authors: Leopoldo Catania

A semiparametric model is developed to track a large number of quantiles of a time series. The model satisfies the condition of non crossing quantiles and the defining property of fixed quantiles. A key feature of the specification is that the updating scheme for time varying quantiles at each probability level is based on the gradient of the check loss function, that forms a martingale difference sequence. Theoretical properties of the proposed model are derived, such as weak stationarity of the quantile process and consistency and asymptotic normality of the estimators of the fixed parameters. The model can be applied for filtering and prediction. We also illustrate a number of possible applications such as: i) semiparametric estimation of dynamic moments of the observables, ii) density prediction, and iii) quantile predictions.

E1259: **Modelling corporate defaults: A Markov-switching Poisson autoregressive model**
Presenter: **Geir Berentsen**, University of Bergen, Norway
Co-authors: Jan Bulla, Antonello Maruotti, Baard Stove

A new model is presented for count time series useful for modelling corporate defaults. The model is an extension of a traditional autoregressive count time series model, where the parameters are allowed to depend on the state of an unobserved Markov chain. The estimated model for US monthly corporate defaults indicates (at least) two regimes and that the so-called contagion effect, that current defaults affect the probability of other firms defaulting in the future, is more present in one of these regimes, even after controlling for financial and economic covariates. The effects of financial and economic covariates are also significantly different in each of the regimes. These results imply that the notion of contagion in the default count process is time-dependent, and thus more dynamic than previously believed.

EG009 Room MAL 354 CONTRIBUTIONS IN SURVIVAL AND RELIABILITY Chair: Mariangela Zenga

E1718: **Kullback-Leibler goodness-of-fit tests for exponential simple step-stress accelerated life testing models**
Presenter: **Anastasia Gaponik**, RWTH Aachen University, Institute of Statistics, Germany
Co-authors: Maria Kateri

Step-stress accelerated life testing (SSALT) experiments are used for assessing the reliability of products in reasonable time. Under a SSALT model, the test units are exposed to stress levels that increase at intermediate time points. For the SSALT model specification, which influences the associated statistical inference procedures, there are three core assumptions to make: the underlying lifetime distributions for each stress level, the model for the joint cumulative distribution function, and the link model that connects the respective lifetime to the stress. To ensure accurate results of reliability prediction, a justification for the chosen assumptions is required. We focus on testing the validity of the first assumption. It is common practice to assume exponential distributed lifetimes for the units under test. In spite of the rich bibliography on goodness-of-fit testing of exponentiality, the literature on goodness-of-fit tests for exponential SSALT models is restricted. We consider a simple SSALT model (i.e. consisting of two stress levels) under the cumulative exposure model and investigate tests based on the Kullback-Leibler information, comparing different entropy estimators. Moreover, via Monte Carlo simulations, the Kullback-Leibler test statistics are compared in terms of power to several common nonparametric goodness-of-fit tests against Weibull and Gamma distributed alternatives.

E1863: **A class of tests for trend for time censored recurrent event data not following a Possion process**
Presenter: **Jan Terje Kvaloy**, University of Stavanger, Norway
Co-authors: Bo Henry Lindqvist

Many tests for trend have been derived under a Poisson process assumption, but it is well known that such tests are not robust against deviations from the Poisson assumptions commonly seen in practice. Several tests for trend in recurrent event data not following a Poisson process are proposed, but these are generally constructed for event-censored data. However, time censored data are more frequently encountered in practice. Our contribution is to present a class of statistical tests for trend in time censored recurrent event data, based on the null hypothesis of a renewal process. The class of tests is constructed by an adaptation of a functional central limit theorem for renewal processes. By this approach a number of tests for time censored recurrent event data can be constructed, including among others a new version of the classical Lewis-Robinson trend test and an Anderson-Darling type test. The latter test turns out to have attractive properties for general use by having good power properties against both monotonic and non-monotonic trends. Extensions to situations with several processes are considered. Properties of the tests are illustrated by simulations and applications to real data.

E1679: **Regular HIV testing: Why it matters and how to measure it**
Presenter: **John Rice**, University of Colorado, Denver, United States
Co-authors: Brent Johnson, Robert Strawderman

Screening for infectious diseases such as HIV is an important public health priority, but traditionally only the rate of screening has received attention. We argue that it is equally important to examine the regularity of screening. Modeling testing events as a renewal process, we show that the mean delay in diagnosis increases quadratically with the coefficient of variation (CV) of the intertest times, a quantity that is inversely proportional to regularity. This result implies that greater regularity in screening is crucial from a public health perspective because delays in diagnosis may lead to worse prognoses and greater likelihood of unaware patients infecting others. In order to estimate this delay in a population of at-risk individuals, we fit regression models for the mean and CV of the intertest time distribution to data in which study participants were asked to recall the time of their most recent HIV test within pre-specified intervals. Finally, we present power analysis of a likelihood ratio test comparing the mean and CV between two groups, concluding that while efficiency is lost with discretization of backward recurrence times, an optimal set of intervals may be chosen such that this loss is minimized.

E1801: **In-sample hazard forecasting based on survival models with operational time applied to non-life reserving**
Presenter: **Stephan Bischofberger**, Cass Business School, United Kingdom

A generalization of the accelerated failure time model allowing the covariate effect to be any positive function of the covariate is introduced. The covariate effect and the baseline hazard rate are estimated nonparametrically via an iterative algorithm. In an application in non-life reserving in actuarial science, the survival time models the development delay of a claim and the covariate effect is often called operational time. Time of underwriting serves as covariate. The estimated hazard rate is a nonparametric alternative to development factors in reserving and is used to forecast outstanding liabilities. Hence, we provide an extension of the chain-ladder framework without the assumption of independence between delay and underwriting.

EP863 Room Macmillan Hall and Crush Hall POSTER SESSION CMSTATISTICS II Chair: Panagiotis Paoullis

E1661: **R package OptimalDesign**
Presenter: **Lenka Filova**, Comenius University in Bratislava, Slovakia
Co-authors: Radoslav Harman

The purpose is to present computational and graphical capabilities of our R package OptimalDesign which provides a toolbox for the computation of D-, A-, I-, and c-efficient exact and approximate designs of experiments on finite domains, for regression models with real-valued, uncorrelated observations. The package fills a gap in presently available functions for experimental design optimization by implementing several competing algorithms based on significantly different principles, including mathematical programming methods and search heuristics. This allows the user to cross-check the quality of the results, and, in many cases, provides efficient design alternatives to choose for the experiment. An important feature of the package is that several implemented procedures allow for multiple linear constraints on the vector of design weights.

E1706: **Estimation of reference curves for fetal weight**
Presenter: **Sandie Ferrigno**, INRIA NANCY / University Nancy Lorraine, France
Co-authors: Myriam Maumy-Bertrand

Reference or standard curves are required in many biomedical problems. Values which lie outside the limits of these reference curves may indicated the presence of disorder. Data are from the French EDEN mother-child cohort. We are studying fetal weight that depends on the gestional age in the second and the third trimester of mother's pregnancy. Some classical parametric methods as LMS method are used to construct these curves. However, they require strong assumptions. So, we also propose to estimate these curves by using nonparametric methods as local polynomial estimation. Their performances are illustrated by comparison with parametric methods on the data set.

E1744: **Modelling the driving speed on expressway ramps**
Presenter: **Jan Elgner**, CDV - Transport Research Centre, Czech Republic
Co-authors: Veronika Rimalova, Eva Fiserova, Jiri Ambros

The question whether there are any causal factors affecting the speed of vehicles in road traffic is of great interest. Identifying such causal effects encounters some difficulties, due to wide range of aspects affecting vehicle drivers, e.g. road traffic volume, road curvature and cross-section characteristics, type of vehicle, road type, weather condition, etc. The example data set was collected from GPS on-board units of floating car fleet on 6 expressway ramps in the city of Brno, Czech Republic for half a year. It contains approximately 400 unique drives on each ramp. The vehicle speed data were recorded every 0.25 seconds, representing a nearly continuous phenomenon, and therefore it is convenient to employ the functional analysis (FDA) techniques. The development of FDA arose from the need for analysing complex data structures, such as curves or surfaces, and many statistical methods have been modified for functional data. In this particular case, the speed profiles are treated as functions of distance. The main interest lies in finding patterns in driving behaviour on the expressway ramps. This can be reached by examining functions derivatives, or by finding relation between speed profile and other influential aspects, either scalar or functional.

E1933: **Analysing administrative data using logistic regression modelling**
Presenter: **Maria de Fatima Salgueiro**, Instituto Universitario de Lisboa (ISCTE-IUL) and Business Research Unit (BRU-IUL), Portugal
Co-authors: Marcel Vieira, Peter W F Smith

A binary logistic regression model was estimated with big real register data, using population and sample values from CadUnico, a Brazilian administrative data source used to select low income families for the anti-poverty Bolsa Familia programme. The target population includes over 27 million families. Samples were selected by alternative probability sampling designs, namely simple and stratified simple random sampling with equal, proportional and optimum allocation in the strata. Different sampling fractions were considered. A binary logistic regression model was estimated to explain the probability of a family receiving the benefit, as a function of 16 covariates. In total, 31 parameters were estimated, and 324 seconds were required to achieve convergence (i5 processor, 16 GB RAM memory). Probability samples of 1 and 5% were selected and the chosen population model was estimated. A forward selection procedure was considered to include covariates in the model. Results suggest that 5% samples are enough to reproduce the odds ratio structure of the chosen population model, especially when simple or stratified simple random sampling with proportional allocation were adopted. Moreover, the adoption of sampling procedures lead to a considerable reduction of computational time (down to 19 seconds for the 5% simple random sample), allowing for a faster modelling decision-making process with a standard personal computer.

E1937: **A comparison of methods to deal with observations below a limit of detection**
Presenter: **Maria Polidoro**, ESTG|P. Porto and CEAUL|FCUL, Portugal

In many applied settings, often due to equipment limitations, measures below a certain limit of detection (LOD) cannot be obtained. Earlier approaches addressed this issue by omitting the observations below the LOD or by imputing them using values such as the LOD/2. However, these strategies can lead to biased estimates of the quantities of interest. Over the last decade or so, several approaches, both within the classical and Bayesian paradigms, have been proposed to properly handle observations that fall below the LOD. The goal is to compare several of these proposed methods via an extensive simulation study.

E1902: **Dimension reduction methods for multilevel neural firing rate data**
Presenter: **Angel Garcia de la Garza**, Columbia University, United States
Co-authors: Jeff Goldsmith

Recent advances have allowed high-resolution observations of firing rates for a collection of individual neurons; these observations can provide insights into patterns of brain activation during the execution of tasks. Our data come from an experiment in which mice performed a reaching motion following an auditory cue, and contain measurements on firing rates from neuron in the motor cortex before and after the cue. We focus on appropriate dimension reduction techniques for this setting, in which sharp increases in firing rates after the cue are expected; we also allow for correlation across neurons in each of the trials, and within a neuron across trials. Initial results suggest differing patterns of activation, perhaps representing the involvement of different motor cortex functions at different times in the reaching motion.

E1448: **The cube transformation of the left truncated normal distribution**
Presenter: **Awa Ogbonnaya Dike**, Akanu Ibiam Federal Polytechnic, Nigeria

The n^th power transformation, a general rule to all power transformations, is used to derive the probability density function (pdf), the mean and variance of the cube transformation of the left truncated normal distribution. This was achieved by substituting $n = 3$ in the general rule. The results obtained are useful in time series modeling where the error component of a multiplicative time series $et > 0$. The real life application of data

on crude oil production in Nigeria between 2009 and 2017, gave the regression equation as $LogStdev = 1.34 - 2.15LogMean$ with $slope = -2.15$ which agrees with the cube transformation as given in the extension of Bartletts transformation table.

E1647: **Asymmetric beta-transformed linear opinion pooling for modeling unbalanced binary data**
Presenter: **Tim Bal**, Ghent University, Belgium
Co-authors: Thierry Marchant

Previous research in opinion pooling used the beta-distribution to transform linearly aggregated probabilities, in order to improve the predictions, in terms of the Brier Score. We propose a similar method, with the main difference that our method does not put any constraint on the shape parameters α and β of the beta-distribution. Our method outperforms previous methods as well as logistic regression as soon as we start dealing with unbalanced data (i.e., when $\mathcal{P}(Y=1) \neq \mathcal{P}(Y=0)$), in terms of the Brier Score and more specifically the calibration. Next to this, we argue that our method is preferred over a recently proposed Bayesian approach since both methods achieve the same results in terms of the Brier Score, but where the Bayesian approach can take up to half an hour, our method needs only a matter of seconds. All methods are compared using simulation studies with the skew-normal distribution and real-life data concerning the validation of diagnosis of ADHD.

E1841: **Multifractal statistics for characterising two-dimensional spatial distribution of population and stores/facilities**
Presenter: **Mariko Ito**, The University of Tokyo, Japan
Co-authors: Takaaki Ohnishi

The spatial distribution of population and stores/facilities is generally heterogeneous. When we investigate such an object in which the density is heterogeneously distributed, multifractal analysis is a good tool to characterize the structure. In multifractal analysis, for each dimension (singularity strength), we derive the fractal dimension (spectrum) of points in which the local fractal dimension around each of those is the singularity strength. We use Japanese 100-meter estimated mesh data from national censuses and corporate telephone directory database teleport with coordinates, as the data of the spatial distribution of population and stores/facilities respectively. We perform multifractal analysis on these data. Following the frequently used derivation, we calculate the spectrum from the q-th generalized dimension, where q is an integer running in a certain region. For two spectrums, for each q, we measure the distance between those coordinates of spectrums derived from q-th generalized dimension. By using this distance, we further define the distance between the two spectrums. We perform clustering of stores/facilities based on this spectrum distance. We discuss what kind of stores/facilities are in the same cluster as the one of population, and its economic background.

E1883: **Generalized mixed model with non-normal density**
Presenter: **Yusuke Saigusa**, Yokohama City University, Japan
Co-authors: Osamu Komori, Shinto Eguchi

Generalized linear mixed model (GLMM) is used frequently in the analysis of non-independent data which includes longitudinal data. In GLMM, the biased estimates and/or poor coverage rates for confidence intervals can arise from the misspecification of the distribution of random effects. We newly propose an extended GLMM with predictor formed by generalized average. The predictor of proposed model has non-normal density whereas the linear predictor of GLMM has normal density when the random effect part has a normal distribution. The relaxed constraint of proposed model would reduce the influence of misspecification on random effects when the distributional assumption is violated. We obtain a computational algorithm for estimating fixed and random effect parameters based on penalized quasi-likelihood, and variance components of random effects based on restricted maximum likelihood in the proposed model. The conditional Akaike information criterion is used for model selection. We give some simulation experiments and example of analysis of longitudinal data from randomized clinical trial for comparing treatments for epileptics.

E0918: **Phylogenetic structures of the environmental niches of phytoplankton: Diatoms vs. dinoflagellates**
Presenter: **Zhi-Ping Mei**, Dalhousie University, Canada
Co-authors: Andrew Irwin, Zoe Finkel

Niche differences among phyla of phytoplankton are well known, but less is known about niche differentiation within phyla. Ecological niches of phytoplankton collected at Gullmar time series station (Sweden) during 1989 through 2014 were modelled with MaxEnt. Phylogenetic principal component analysis and variance decomposition of niches across the phylogenetic tree were conducted to study phylogenetic signal of niche differentiation of diatoms and dinoflagellates, determining the extent to which niches are conserved across the phylogenetic tree of the two functional groups. There is clear distinction in niches between diatoms and dinoflagellates, with diatoms in lower temperature, shorter daylength, weaker water column stratification, and higher nutrient niches than dinoflagellates, with the exception that Gymnodiniales of dinoflagellates share similar range of ecological niches with diatoms. Nutrient and physics niches tend to be conserved within two functional groups of diatoms and dinoflagellates. Within each group, different niche variables diverged at different nodes at the lower taxonomic levels of the phylogeny. On average, 15 - 40% of the variation in niche means across species is determined by phylogenetic branches older than subclasses for diatoms, or orders for dinoflagellates. Niche divergence at different nodes of the phylogenetic tree is consistent with the evolution of ocean environment throughout the evolutionary history of diatoms and dinoflagellates.

E1991: **Finding significant and insightful relationships between biotic and abiotic factors in Atlantic river corridors**
Presenter: **Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Maria Fernandez-Garcia, Mauro Sanna, Gil Gonzalez-Rodriguez, Jorge Marquinez

River corridors are subject to numerous pressures such as invasive species, plant diseases, intensification of uses, artificial structures and activities, etc. These pressures affect their functions as reservoirs of diversity of flora and fauna species, flood peak damping, hydrological cycle regulators and ecosystem service providers. Edaphological, geomorphological and vegetation characteristics have been mapped for an initial diagnosis of the conservation status in several rivers in the NW of Iberian Peninsula. We examine the correlation among more than 40 biotic and abiotic factors in order to study their effects on restoration activities and improve the integral management of river corridors.

E2021: **A Bayesian approach for estimating the causal effects using sparse invalid instrumental variables**
Presenter: **Shunsuke Horii**, Waseda University, Japan

The estimation of the causal effect of a potentially endogenous treatment on some outcome is studied. One of the ways to handle the endogeneity is to use the additional variables called instrumental variables. Instrumental variables are correlated with the endogenous treatment variable, but they do not have direct effects on the outcome. The latter condition is called excludability. It is known that the excludability condition is not testable, and sometimes it may be violated. We employ a Bayesian estimation in models that consider the endogeneity and shrinkage prior to the instrumental variables which may not satisfy the excludability conditions. The assumption that the excludability condition of the instrumental variables may be violated is modeled by assuming horseshoe prior to the regression coefficients on the regression from the instrumental variables to the objective variable. We show that the Bayesian inference algorithm can estimate the target causal effect correctly through simulation experiments.

E2023: **Influenza epidemics are associated with temperature variability in Central Europe**
Presenter: **Jan Kysely**, Institute of Atmospheric Physics AS CR, Czech Republic
Co-authors: Hana Hanzlikova, Ales Urban, Jan Kyncl

Influenza and acute respiratory infections (ARI) show seasonal patterns with winter incidence peaks and significantly contribute to excess winter mortality. A recent research suggests that influenza virus survival, transmission and seasonality can be modulated by weather conditions, such as periods of cold and dry air. We examined the relationships between environmental factors, epidemics of influenza/ARI and all-cause mortality in the population of the Czech Republic over 1982-2015. Epidemics were defined from weekly data on influenza/ARI morbidity, and differences between

prevailing virus types were considered. The all-cause mortality data were controlled for long-term trends, seasonality, and the weekly cycle. The results showed that the seasonal onset of influenza/ARI epidemics was associated with temperature variability. The epidemics were typically preceded by 3-week periods of temperature fluctuations and prevailing westerly flow, suggesting frequent passages of air masses with humid air over Central Europe. The onset of severe epidemics with dominant H3N2 virus type was accompanied by pronounced changes in meteorological conditions, namely reduced westerly flow, intense cooling and low temperatures persisting several weeks. The relatively mild ambient temperatures before epidemics favour virus transmission and spreading in population while low temperatures during epidemics may enhance virus infectivity and prolong its survival.

E2028: **Model-based clustering for large scale data with a massive null group**
Presenter: **Soohyun Ahn**, Ajou University, Korea, South

A new clustering method for large scale data with a massive null group is proposed and called self-semi-supervised clustering. Self-semi-supervised clustering is a two-stage procedure: pre-select a part of "null" group from the data in the first stage and apply semi-supervised clustering to the rest of the data in the second stage, allowing them to be assigned to the null group. We evaluate the performance of the proposed method using a simulation study and demonstrate the method in the analysis of time course gene expression data from a longitudinal study of Influenza A virus infection.

E2036: **Variable selection and prediction in logistic regression with incomplete data**
Presenter: **Rong Xing**, Shanghai University of International Business and Economics, China
Co-authors: YunXiang Cao

Multiple imputation random lasso (MIRL) method is an extension of the random lasso to combine penalized regression techniques with multiple imputation techniques. This paper aims to apply MIRL to logistic regressions in order to deal with classification problem with incomplete high-dimensional predictors. The missing completely at random pattern is considered. Extensive simulation studies are conducted to compare MIRL-logistic with its several alternatives. The result shows that MIRL-logistic has an improved accuracy on both outcome prediction and variable selection performance in high-dimensional scenarios. Especially, the proposed method performs well when the correlation among the predictors is high.

E1888: **Improving the probability weighted moment to estimate the shape Pareto model**
Presenter: **Ayana Mateus**, NOVA.ID.FCT - Universidade Nova de Lisboa, Portugal
Co-authors: Frederico Caeiro

The Pareto distribution was first introduced as a model for large incomes and nowadays has been extensively used for modelling events in fields such as bibliometrics, demography, insurance, finance, risk management, biology and astronomy. We propose a consistent estimator for any positive shape parameter of the Pareto distribution. This objective is achieved through a modification of the probability weighted moments method. We also compare, through a Monte Carlo simulation study, the finite sample performance of the proposed estimator, in terms of the mean value and root mean square error, with the most usual estimators from the literature.

E2014: **Extremal index blocks estimator: Another approach**
Presenter: **Dora Prata Gomes**, NOVA.ID.FCT FCT-UNL, Portugal
Co-authors: Manuela Neves

The main objective of Statistics of Extremes is the estimation of probabilities and parameters related to rare events. When extending the analysis of the limiting behaviour of the extreme values from independent and identically distributed sequences to stationary sequences a key parameter appears, the extremal index, whose accurate estimation is not easy. The extremal index measures the degree of local dependence in the extremes of a stationary process. We focus on the estimation of the extremal index using blocks estimators. Blocks estimators can be constructed by using disjoint blocks or sliding blocks but both blocks estimators require the choice of a threshold and a block length. Some criteria have appeared for the choice of those nuisance parameters. The objective is to revisit another block estimation procedure that only depends on the block length, although some conditions on the underlying process need to be verified. The associated estimator presents nice asymptotic properties, and for finite samples is illustrated a stability criterion for choosing the block length and then obtaining the extremal index estimate. A simulation study has been performed to illustrate the behaviour of the blocks estimator and an application to real data is presented.

E2035: **Selecting an optimal cutpoint in Cox proportional hazards models with several covariates**
Presenter: **Woojoo Lee**, Inha University, Korea, South

In survival analysis, a continuous covariate sometimes needs to be transformed to a binary variable to enhance interpretation of regression coefficients. For doing this, the key problem is to find an optimal cutoff and assess the statistical significance of the transformed binary variable correctly. A naive approach is using the cutoff giving the most significant p-value for the binary variable, but this shows highly distorted type I error. In order to overcome such problem, two novel testing methods were developed based on Brownian or Brownian bridge process. Although these methods assumed that there is only one continuous covariate in Cox proportional hazards model, which is often violated in practice, they are currently employed as standard methods even when there are several covariates. We investigate the performance of the two methods when there are several covariates in Cox proportional hazards model. Our numerical study shows that the two testing methods may not provide an optimal cutoff and suffer from distorted type I error. To mitigate these problems, we adopt a more recent testing method that can take into account other covariates in the Cox model. The numerical study shows that the recent testing method controls type error well at the required nominal level and shows smaller mean squared error of cutoff estimator than the two methods developed for only one continuous covariate case.

CI018 Room Chancellor's Hall STATIONARITY AND CAUSALITY OF TIME SERIES Chair: Josu Arteche

C0176: **A cointegrated model allowing for different fractional orders**
Presenter: **Morten Nielsen**, Queen's University, Canada
Co-authors: Soren Johansen

Earlier work on the FCVAR model is generalized to allow each observed variable to have its own fractional integration order. We discuss cointegration, representation, and statistical inference in the model, and derive the relevant asymptotic theory. In particular, the asymptotic distribution of (linear combinations of) the maximum likelihood estimators of the fractional orders can be superconsistent and mixed Gaussian depending on the connectedness of the graph of the identified beta vectors. In other cases, the estimators of the fractional orders are Gaussian. The remaining estimators and test statistics retain their asymptotic properties from the more standard FCVAR model.

C0177: **Powerful self-normalizing tests for stationarity**
Presenter: **Uwe Hassler**, Goethe University Frankfurt, Germany
Co-authors: Mehdi Hosseinkouchack

A family of tests for stationarity against a unit root is proposed. It builds on the Karhunen-Loeve expansions of the limiting CUSUM processes under the null hypothesis and local alternatives. The test statistic becomes a ratio of quadratic forms of q weighted sums such that the nuisance long-run variance cancels asymptotically without having to be estimated. Critical values can be calculated by standard numerical means. Monte Carlo experiments show that q may not be too large in finite samples to obtain a test with correct size under the null. At the same time our test is more powerful than classical competitors that are not self-normalizing.

C0178: **Identification of possibly nonfundamental Structural VARMA models using higher order moments**
Presenter: **Carlos Velasco**, Universidad Carlos III de Madrid, Spain

A frequency domain criterion is introduced to identify the parameters of, possibly noncausal and/or noninvertible, structural vector autoregressive moving average (SVARMA) models. We use information from higher order moments to achieve identification on the location of the roots of the VAR and VMA matrix polynomials for possibly non-fundamental non-Gaussian vector time series. This information also provides identification on the rotation of the model errors leading to the structural innovations up to sign and permutation. We develop general representations of the higher order spectral density arrays of vector linear processes and describe sufficient conditions for the parameter identification that rely on both sufficiently rich (linear) dynamics and the independence component structure of the vector of linear innovations. We generalize previous univariate analysis to develop efficient estimates exploiting linear and higher order dynamics.

CO693 Room Bloomsbury FINANCIAL ECONOMETRICS: HIGH-FREQUENCY OPTION DATA RESEARCH Chair: Ingmar Nolte

C0228: **Aggregate asymmetry in idiosyncratic jump risk**
Presenter: **Viktor Todorov**, Northwestern University, United States

The structure and pricing of idiosyncratic jumps, i.e., jumps in asset prices that occur outside market-wide jump events, are studied. Using options on individual stocks and the market index that are close to expiration as well as local estimates of market betas from returns on the underlying assets, we estimate nonparametrically the asymmetry in the risk-neutral expected idiosyncratic variation, i.e., the difference in variation due to negative and positive returns, which asymptotically is solely attributed to jumps. We derive a feasible Central Limit Theorem that allows us to quantify precision in the estimation, with the limiting distribution being mixed Gaussian. We find strong empirical evidence for aggregate asymmetry in idiosyncratic risk which shows that such risk clusters cross-sectionally. The results reveal the existence and non-trivial pricing of aggregate downside tail risk in stocks during market-neutral systematic events as well as a negative skew in the cross-sectional return distribution during such episodes.

C0244: **Spatial dependence in option observation errors**
Presenter: **Torben G Andersen**, Kellogg School, Northwestern University, United States
Co-authors: Viktor Todorov, Nicola Fusari, Rasmus Varneskov

A nonparametric test is developed for deciding whether the observation error in option panels has spatial dependence. The option panel consists of options written on an underlying asset with different strikes and times to maturity. The asymptotic setup is of infill type: the mesh of the strike grids of the observed options shrinks asymptotically to zero while keeping the set of observation times and maturities fixed. We propose a Ljung-Box type test for testing the null hypothesis of no spatial dependence in the observation error. The test makes use of the smoothness of the true (unobserved) option price as a function of its strike and is robust to presence of heteroskedasticity of unknown form in the observation error. A Monte Carlo study shows good finite sample properties of the developed testing procedure and empirical application to S&P 500 index option data reveals mild spatial dependence in the observation error which has declined over time.

C0381: **A descriptive study of the high-frequency trade and quote option data from OPRA**
Presenter: **Manh Cuong Pham**, Lancaster University, United Kingdom
Co-authors: Ilya Archakov, Leon Grund, Nikolaus Hautsch, Sergey Nasekin, Ingmar Nolte, Stephen Taylor

A guide to high frequency option trade and quote data disseminated by the Options Price Reporting Authority (OPRA) is provided. First, we present a comprehensive overview of the fragmented U.S. option market, including details on market regulation and the trading processes for all 15 constituent option exchanges. Then, we review the general structure of the OPRA dataset and present a thorough empirical description of the observed option trades and quotes for a selected sample of underlying assets that contains more than 25 billion records. We outline several types of irregular observations and provide recommendations for data filtering and cleaning. Finally, we illustrate the usefulness of the high frequency option data with two empirical applications: option-implied variance estimation and risk-neutral density estimation. Both applications highlight the superior information content of the high frequency OPRA data to the widely used end-of-day OptionMetrics data.

C0819: **A multivariate realized GARCH model**
Presenter: **Ilya Archakov**, University of Vienna, Austria
Co-authors: Peter Hansen, Asger Lunde

A novel class of multivariate Realized GARCH models is proposed which is based on a convenient parametrization of the correlation matrix. The correlation matrix is characterized by a vector that can vary freely in the real space. A more parsimonious structure is often desired in practice, in particularly in high dimensional systems, and the model facilitates a simple and intuitive dimension reduction of the parametrized correlation matrix. We apply the model to returns of nine assets and demonstrate the dimension reduction by adopting a natural block correlation structure.

CO238 Room G11 ADVANCES IN EXACT AND APPROXIMATE BAYESIAN COMPUTATION Chair: Robert Kohn

C0862: **Automated sensitivity computations for Bayesian Markov chain Monte Carlo inference: A new approach**
Presenter: **Dan Zhu**, Monash University, Australia
Co-authors: Liana Jacobi

An efficient numerical approach is introduced to implement a comprehensive sensitivity analysis of MCMC output with respect all input parameters, i.e. prior hyper-parameters and chain starting values. Building on recent developments of automatic differentiation (AD) in the classical simulation setting, we develop an AD MCMC scheme that is applicable to MCMC algorithms composed of both continuous and dis-continuous high-dimensional mappings. It enables the computation the of sensitivities based on exact (up to computer floating point error) first-order derivatives of MCMC draws alongside the estimation algorithm. The new approach makes it computationally feasible to (i) undertake a complete local robustness analysis of a wide range of posterior output with respect to all prior input parameters; and (ii) assess algorithm performance, in particular convergence behaviour, via the evolution of starting value sensitivities. We discuss a wide range of prior robustness measures, including new measures relating to overall model robustness and quantile robustness. In addition, convergence diagnostics based on the evolution of overall starting value sensitivities are introduced. Performance and applications of the method are illustrated in simulated and real data examples.

C1169: **Recent developments in data subsampling for large-scale Bayesian inference**
Presenter: **Mattias Villani**, Linkoping University, Sweden
Co-authors: Matias Quiroz, Robert Kohn, Minh-Ngoc Tran, Doan Khue Dung Dang

Hamiltonian Monte Carlo (HMC) is an increasingly popular simulation algorithm for Bayesian inference which has proven to be especially suitable in high-dimensional problems. A drawback of HMC is that it requires a large number of evaluations of the posterior and its gradient, which can be computationally costly, particularly in problems with large datasets. Results on accelerating HMC by data subsampling and how to optimally tune the algorithm are presented. Extensions to dependent data are also discussed.

C1196: **Variational Bayes on manifolds**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Co-authors: Dang Nguyen, Duy Nguyen

Increasingly complicated models in modern statistics have called for more efficient Bayesian estimation methods. A Variational Bayes algorithm is developed that exploits both the information geometry of the manifold of probability distribution functions and the manifold structure of the

variational parameters. The information geometry of the manifold of probability distributions results in the natural gradient which is the steepest ascent on this manifold. Utilising the manifold structure of the variational parameters leads to an efficient non-linear optimization technique that takes into account the constraint structure of the parameter space. The convergence of the proposed algorithm is theoretically guaranteed and its performance is tested on several challenging examples including deep neural networks.

C1981: **Fast variational approximation for multivariate factor stochastic volatility models**
Presenter: **David Gunawan**, University of New South Wales, Australia
Co-authors: David Nott, Robert Kohn, Matias Quiroz

Estimating and predicting the densities and time-varying correlation matrices for high dimensional time series are important and an active area of research. One of the main challenges is the curse of dimensionality, where the number of parameters in the covariance matrix grows quadratically with the number of time series considered. The factor SV model is an important multivariate time series model for parsimoniously modeling a vector of time series. It imposes a much lower dimensional latent factors that are allowed to exhibit stochastic volatility and govern the comovement of the time series over time. One of the main challenges of estimating this model is that it has both large number of latent states and large number of parameters. This model is usually estimated by using MCMC or particle MCMC method, which can be slow for high dimensional and long time series. Fast sequential and batch variational estimation methods are proposed to approximate the posterior distribution of the states and parameters in a multivariate factor stochastic volatility (SV) model and obtain one-step and multiple step-ahead forecast distribution. We apply our method to simulated and real datasets.

CO797 Room G3 THE THEORY, APPLICATIONS AND COMPUTING OF INDICATOR SATURATION Chair: James Reade

C0518: **A simple robust procedure in instrumental variables regression**
Presenter: **Xiyu Jiao**, University of Oxford, United Kingdom

Due to the frequent concern that outliers may invalidate the empirical findings, in practical applications of instrumental variables regression the common practice is to first run ordinary two stage least squares and remove observations with residuals beyond a chosen cut-off value that classifies outliers. 2SLS is subsequently re-calculated with non-outlying observations, and this procedure can be iterated until robust results are obtained. We analyze this simple robust algorithm asymptotically, then provide consistent estimation and valid inferential procedures for practical implementation given the cut-off value. Moreover, we provide asymptotic theory for setting the cut-off, which is chosen to control the gauge (proportion of outliers wrongly discovered). Asymptotics are derived under the null hypothesis that there is no contamination in the cross-sectional i.i.d. data. The established weak convergence result, involving empirical processes and fixed points, provides a starting point for statistical tests that formalize robustness checks on the difference between ordinary and robust 2SLS.

C0529: **Modelling Australian electricity price using indicator saturation**
Presenter: **Shixuan Wang**, University of Reading, United Kingdom
Co-authors: James Reade

Indicator saturation on electricity price series from the National Electricity Market (NEM) in Australia is employed in order to model the stylized facts of electricity prices, which include extreme spikes, seasonality, level-shifts, and autocorrelation. Standard modelling techniques in the literature to cope with these characteristics tend to focus on regime-switching models, which is constrained by the limited number of different, uncertain structural breaks, and the numerical estimation which could be possibly not converged. In particular, we develop an iterated procedure to capture outliers and shifts of differing magnitudes and economic importance. Based on a range of model evaluation tools, we find that indicator saturation method outperforms the regime-switching models with various settings. In additional to the statistical superiority, we further link our results of significant spikes, level-shifts, and trends to the policy changes in the NEM and provide recommendations for the development for the electricity markets in Australia.

C0575: **Modelling non-stationary 'big data'**
Presenter: **Jennifer Castle**, Oxford University, United Kingdom
Co-authors: Jurgen Doornik, David Hendry

Seeking substantive relationships among vast numbers of spurious connections when modelling Big Data requires an appropriate approach. Big Data are useful if they can increase the probability that the data generation process (DGP) is nested in the postulated model, increase the power of specification and mis-specification tests, and yet do not raise the chances of adventitious significance. Simply choosing the best-fitting equation or trying hundreds of empirical fits and selecting a preferred one–perhaps contradicted by others that go unreported–is not going to lead to a useful outcome. A crucial issue addressed in this paper is that wide-sense non-stationarity (cointegration and location shifts) must be taken into account if statistical modelling by mining Big Data is to be productive. A factor approach to identifying cointegrating relationships is investigated. Moreover, important computational problems must be resolved given the huge numbers of possible models to be selected over.

C1103: **Asymptotic properties of the gauge of step-indicator saturation**
Presenter: **Matthias Qian**, Oxford University, United Kingdom
Co-authors: Bent Nielsen

The aim is investigated the asymptotic properties of the gauge of Step-indicator Saturation which is an algorithm to handle unmodelled location shifts in time series. The gauge is the frequency of falsely retained step-indicators when the data generating process has no shifts. We present asymptotic convergence and distribution results of the gauge of the algorithm. The proofs rely on empirical process results of temporal differences of residuals. Our results offer an asymptotic justification to use the gauge in choosing the only tuning parameter of this statistical procedure.

CO386 Room G4 ROBUST MODELS IN THE TIME AND FREQUENCY DOMAINS FOR HIGH DIMENSIONAL DATA Chair: Pascal Bondon

C0736: **Iterative robust hypothesis testing for change-points detection and application to SAR change detection**
Presenter: **Ammar Mian**, CentraleSupelec SONDRA, France
Co-authors: Guillaume Ginolhac, Jean-Philippe Ovarlez, Abdourahmane Atto

The problem of detecting change-points in a time series of multivariate sets of vectors is considered by exploiting covariance homogeneity testing schemes. Notably, we propose to consider the problem under a robust framework by assuming Complex Elliptically Symmetric distribution models for which we tailor detection tests. The robustness of this approach, compared to one based on a Gaussian assumption, will be considered through theoretical and experimental analysis. Then, based on these developments, we consider an iterative algorithm to determine the points of change in a time series of data. Finally, an application to the analysis of changes in Synthetic Aperture Radar time series will be proposed to demonstrate the interest of the robust framework in real-world applications.

C0738: **Median of means estimation for high dimensional time series**
Presenter: **Stephane Chretien**, NPL, United Kingdom

Linear dynamical systems with generalized linear observation models are considered. These models are traditionally estimated using maximum likelihood or moment based methods. Both approaches need in particular to tackle the joint estimation of the latent states which model the dependencies, and the transition parameters which model the dynamics. Recent work proposes to address the estimation problem using optimised

matrix factorisation. Our contribution extends the matrix factorisation approach to the setting where outliers may be present, using Median of Means (MoM)-based robustification.

C1473: **Robust singular spectrum analysis: Methodology and application**
Presenter: **Paulo Canas Rodrigues**, Federal University of Bahia, Brazil

Singular Spectrum Analysis (SSA) is a powerful and widely used non-parametric method to analyze and forecast time series. Although SSA has proven to outperform traditional parametric methods for model fit and model forecasting, one of the steps of the SSA algorithm is the singular value decomposition (SVD) of the trajectory matrix, which is very sensitive to the presence of outliers because it uses the L2 norm optimization. Therefore, the presence of outlying observations has a significant impact on the SSA reconstruction and forecasts. The main aim is to introduce several robust alternatives to SSA, where the SVD is replaced by robust SVD and robust PCA alternatives. The SSA and the six robust SSA alternatives are compared in terms of model fit and model forecasting via Monte Carlo simulations based on synthetic and real data, considering several contamination scenarios.

C1215: **A robust estimation and testing of the cointegration order based on the frequency domain**
Presenter: **Igor Souza**, Universidade Federal de Minas Gerais, Brazil
Co-authors: Valderio Anselmo Reisen, Pascal Bondon, Glaura Franco

The aim is to estimate the degree of cointegration in bivariate series. A test statistic for the non-cointegration based on the determinant of the spectral density matrix for the frequencies close to zero by using periodogram based on M-regression method with Huber function. Series are assumed to be $I(d)$, $0 < d < 1$, with parameter d supposed to be known. In this context, the order of integration of the error series is $I(d-b)$, $b \in [0,d]$. The proposed estimator for b is obtained by performing a regression of logged determinant on a set of logged Fourier frequencies. Under the null hypothesis of non-cointegration, the expressions for the bias and variance of the estimator are derived and consistency properties were also obtained. The asymptotic normality of the estimator, under Gaussian and non-Gaussian innovations, were also established. Performance is investigated using Monte Carlo simulations under two scenarios: series uncontaminated and contaminated with additive outliers.

CO400 Room G5 ADVANCES IN FINANCIAL MODELLING Chair: Genaro Sucarrat

C1056: **Liquidity tail risk in the wake of the financial crisis: Evidence from the U.S. stock market**
Presenter: **Debbie Dupuis**, HEC Montreal, Canada
Co-authors: Luca Trapin

Since the last financial crisis, market participants have perceived a deterioration in market liquidity as a consequence of the tightening regulatory constraints. Inspecting common liquidity metrics, recent research in finance however does not find any substantial reduction in liquidity compared to the pre-crisis levels. Going beyond the level of liquidity, the aim is to investigate liquidity tail risk, studying possible changes in the likelihood of extreme illiquidity events over the post-crisis period. To do that, we define a novel state-space extreme value model and build a robust score-driven filter and smoother of the latent states. Fitting the model to several highly liquid stocks in the S&P500 reveals that tail liquidity risk has increased in recent years. While we find insufficient evidence that this increase is solely attributable to a structural change in the dynamics of liquidity provision due to the post-crisis regulatory restructuring, the substantial increase in liquidity tail risk should prompt policymakers and market participants to take mitigating action.

C1322: **Testing asset pricing models on the cryptocurrency market**
Presenter: **Francesco Violante**, ENSAE ParisTech, France
Co-authors: Stefano Grassi

The purpose is to test three and a five-factor models capturing the size, value, momentum and short term reversal in the cross-section of cryptocurrencies returns to examine if these are sufficient to capture market-wide sources of risk. Adapted versions of the Fama-French SMB and HML long-short value-weighted portfolios meaningful to the cryptocurrency market are created using the universe of cryptocurrencies available for the period 2015 to 2019. Inspired by the Cholesky GARCH model, we develop a multivariate conditional beta model based on a block LDU decomposition of the conditional covariance matrix of the system including factors and asset returns. The model allows estimating in a single pass the asset-specific exposure to the risk factors and the premium associated with such exposure. The asset pricing models are tested on the cross-section of 25 value-weighted portfolio sorts by size and network value to transactions ratio.

C0838: **Discrete time nonlinear diffusion models**
Presenter: **Rickard Sandberg**, Stockholm School of Economics, Sweden

A new class Smooth Transition (STR) discrete time diffusion process is considered where the drift component is allowed not only to have a regime-switching type of behavior, but also a time-varying behavior. Such nonlinear features are simply motivated by the fact that the drift of the return series (say) depends on the phases of the business cycle (the regime-switching behavior) as well as the business cycle may be subject to structural changes over time (the time-varying behavior). Simple tests to identify such nonlinearities are derived. Moreover, Quasi Maximum Likelihood Estimation (QMLE) techniques to estimate the proposed discrete time STR diffusion models are discussed. In an application to S&P500 and NYSE data, we find, using our tests, overwhelming evidence in favour of a drift component that is subject to both types of nonlinearities. For an example return series, the QMLE techniques are demonstrated.

C1231: **The fractionally integrated log-GARCH model when the conditional density is unknown**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway

Financial time-series are frequently characterised by strong persistence. This can be due to "long memory" or structural breaks, or both. Recently, Fractionally Integrated log-GARCH (FI-log-GARCH) models have successfully been proposed as a model of such features. A drawback with these contributions, however, is that they rely on the conditional density being known. We propose estimation procedures that do not rely on the density being known. Monte Carlo simulations verify the usefulness of the procedures, and an empirical application provides an illustration.

CO208 Room Gordon SEMI- AND NONPARAMETRIC REGRESSION FOR TIME SERIES AND PANEL DATA II Chair: Harry Haupt

C1512: **Directed local testing in functional linear regression**
Presenter: **Christoph Rust**, University of Regensburg, Germany

A directed testing procedure is developed for the coefficient function of the Functional Linear Model (FLM) with a scalar response. The suggested procedure provides local interpretability of the regression function in the sense that it allows us to infer from the data a threshold up to which the random curves have a significant influence on the scalar outcome variable. The test is applicable in situations where one can assume that the functional predictors influence the outcome variable only on a sub-interval of the domain. Such situations regularly occur for analyzing brain-image data but also in regional economics. Theoretical results are established, and their finite sample performance is assessed by means of a simulation study and shows applicability of the suggested testing procedure with a real data application.

C0789: **Testing for multiple structural breaks in multivariate long memory time series**
Presenter: **Kai Wenger**, Institute of Statistics, Germany
Co-authors: Philipp Sibbertsen, Simon Wingert

Estimation and testing of multiple unknown breaks in multivariate long-memory time series is considered. We propose a likelihood ratio based

approach for estimating breaks in the mean and the covariance of a system of long-memory time series. The limiting distribution of these estimates as well as consistency of the estimators is derived. A test to determine the unknown number of break points is given based on sequential testing on the regression residuals. A Monte Carlo exercise shows the finite sample performance of our method. We illustrate the usefulness of our test by analysing two real data sets.

C1509: **Prediction intervals in quantile regression**
Presenter: **Ida Bauer**, University of Passau, Germany
Co-authors: Harry Haupt, Markus Fritsch

A simulation approach is exploited to investigate prediction intervals in different settings for (generalized) least squares and quantile regression. It provides an extensive and structured overview of different approaches to quantile-based prediction intervals described in the literature and tries a fair comparison to suitable (case-specific) alternatives. We start with a case that complies with all requirements for accurate LS prediction intervals and subsequently relax distributional assumptions. This should allow contrasting strengths and potential drawbacks of each and every method. Evaluation of the accuracy of the intervals is performed by comparing empirical coverage levels as well as taking a look at interval width and percentage of overlap. Further issues to be discussed are whether the fact that regression quantiles used for interval construction are point estimates and thus subject to variation themselves has been recognized in the literature so far and —if so— whether the latter provides a remedy. We also touch upon the topic of quantile crossing, which is a particular issue especially in the context of prediction intervals.

C1451: **Nonlinear quantile regression**
Presenter: **Harry Haupt**, University of Passau, Germany

The nonlinear quantile regression model is studied under heteroskdeasticity and local weak dependence. Departing from recent contributions to the literature we focus on the discussion of CLT by contrasting the assumptions of frequently used Theorems to a simpler CLT, which relies on recent contributions to the theory of weak dependence concepts beyond mixing and near epoch dependence.

CO222 Room Montague REGIME SWITCHING, FILTERING, AND PORTFOLIO OPTIMIZATION Chair: Joern Sass

C0444: **When does portfolio compression reduce systemic risk?**
Presenter: **Luitgard Veraart**, London School of Economics, United Kingdom

The consequences of portfolio compression on systemic risk are analyzed. Portfolio compression is a post-trading netting mechanism that reduces gross positions while keeping net positions unchanged and it is part of the financial legislation in the US (Dodd-Frank Act) and in Europe (European Market Infrastructure Regulation). We show that the recovery rate in case of default plays a significant role in determining whether portfolio compression is potentially beneficial. If recovery rates of defaulting nodes are zero then compression weakly reduces systemic risk. We also provide a necessary condition under which compression strongly reduces systemic risk. If recovery rates are positive we show that whether compression is potentially beneficial or harmful for individual institutions does not just depend on the network itself but on quantities outside the network as well. In particular we show that portfolio compression can have negative effects both for institutions that are part of the compression cycle and for those that are not. Furthermore, we show that while a given conservative compression might be beneficial for some shocks it might be detrimental for others. In particular, the distribution of the shock over the network matters and not just its size.

C0740: **Robust utility maximization in continuous-time financial markets**
Presenter: **Dorothee Westphal**, TU Kaiserslautern, Germany
Co-authors: Joern Sass

Model uncertainty is a challenge that is inherent in many applications of mathematical models in various areas, for instance in mathematical finance and stochastic control. Robust strategies, i.e., strategies that are less vulnerable to the specific choice of the model, are determined by solving worst-case optimization problems. We study utility maximization problems in continuous-time financial markets with uncertainty about the drift and with a constraint on the admissible strategies that prevents a pure bond investment. The drift uncertainty is taken into account by maximizing the worst-case expected utility given that the drift takes values within some uncertainty set. This set is usually motivated by parameter estimations. For a specific choice of uncertainty sets we give an explicit representation of the optimal strategy and prove a minimax theorem. We then show how uncertainty sets can be defined based on filtering techniques and demonstrate that investors need to account for model uncertainty by choosing a robust strategy instead of relying on drift estimations only.

C0772: **Estimating a time-varying parameter model with shrinkage for the Standard&Poor's 500 index**
Presenter: **Borys Koval**, Vienna University of Economics and Business, Austria
Co-authors: Leopold Soegner, Sylvia Fruehwirth-Schnatter

Time-varying parameter models (TVP) are used to investigate in-sample and out-of-sample predictability for monthly returns of the Standard&Poor's 500 index (S&P 500). We consider unrestricted TVP model with a discount factor for the variance process. For the restricted TVP model, we follow the recently developed Bayesian methods that allow for shrinkage for time-varying parameter models. This is attained by applying hierarchical double-Gamma shrinkage priors on the process variances to automatically shrink the time-varying coefficients to static ones. In addition, we differentiate between the significant and insignificant coefficients if the model is overfitted. Both models are tested using simulated data and real market data. Furthermore, we investigate the sensitivity of the estimation approach and the time span used to evaluate the model. To evaluate one-step-ahead predictive densities, Kalman mixture approximations were applied.

C0806: **A multi-asset worst-case approach for optimal portfolios facing crash scenarios**
Presenter: **Elisabeth Leoff**, Fraunhofer ITWM, Germany
Co-authors: Ralf Korn

A worst-case portfolio approach is generalized where the risky asset may have a jump downward at any time, to a multi-asset setting. The non-uniqueness of indifference strategies results in a much more complicated portfolio optimization problem as in the single risky asset framework. To determine the worst-case optimal portfolio processes we develop two new approaches, a Lagrangian multiplier approach in the log-utility case and a combined constrained HJB equation and indifference strategy approach for dealing with power-utility functions. Various examples illustrate remarkable effects and differences compared to the single risky asset setting, in particular the possibility for using some stocks for crash hedging and thereby allowing stock investment possibilities that are not present in the single-stock case.

CO242 Room Woburn MACRO-FINANCE APPLICATIONS Chair: Scott Brave

C0877: **Intermediary asset pricing: Empirical evidence revisited**
Presenter: **Cesare Robotti**, University of Warwick, United Kingdom
Co-authors: Nikolay Gospodinov

New asymptotic tests are proposed on individual asset pricing errors in two-pass cross-sectional regressions. In particular, we show how the individual pricing errors from alternative pricing models can be formally compared in the presence of potential global model misspecification. This complements the large existing literature based on the comparison of aggregate measures of pricing errors. Empirically, we revisit numerous studies on intermediary asset pricing and show that the apparently good performance of leverage-augmented beta-pricing models is due to ignoring

model misspecification and the presence of spurious factors in the analysis. Finally, we provide guidance for reliable inference on risk premia and model performance in linear beta-pricing models.

C1419: **Getting in all the cracks: Monetary policy and indicators of financial stability**
Presenter: **Andrea Ajello**, Board of Governors of the Federal Reserve System, United States

The purpose is to study the effect of monetary policy surprises on aggregate economic activity, price dynamics, and a wide array of measures of financial vulnerability. We do so by building a monetary proxy dynamic factor model (proxy DFM), in which the policy rate, prices, and aggregate activity dynamics interact with measures of asset valuation pressure, and indicators of vulnerability stemming from the financial and non-financial sectors. We find that monetary policy shocks have pervasive effects on financial vulnerability. On the wake of a surprise monetary policy tightening, asset valuations drop on impact and credit standards tighten in the short-to-medium run. While aggregate activity contracts and inflation pressure subsides, measures of debt sustainability and leverage of risky borrowers deteriorate.

C0466: **The network origins of approximate factor models**
Presenter: **Andrew Butters**, Indiana University, United States
Co-authors: Scott Brave, David Kelley

The aim is to investigate the large sample properties of the approximate factor model from within a network model framework. After outlining structural network models that exhibit an approximate factor model reduced form, we propose an alternative estimation strategy and provide normal approximation limit theorems for inference. We also provide Monte-Carlo evidence of the efficiency of the proposed estimator in finite samples relative to benchmark estimators (e.g. principal component analysis). Finally, we apply our estimator to an application involving a large cross-section of economic activity indicators.

C0405: **Optimal debt dynamics, issuance costs, and commitment**
Presenter: **Luca Benzoni**, Federal Reserve Bank of Chicago, United States
Co-authors: Lorenzo Garlappi, Robert Goldstein, Chao Ying

Optimal dynamic capital structure policy in the presence of fixed issuance costs is investigated. For each level of issuance costs, including zero, we identify the global-optimal debt policy. Commitment to this optimal policy is credible if debt holders threaten to punish any deviation by forever pricing debt according to the no-commitment policy. However, commitment to debt repurchases is not credible once realistic equity issuance costs are accounted for, providing one explanation for why firms are unable to issue risk-free debt. When calibrated to realistic issuance costs, the no-commitment policy generates almost as much tax benefits to debt as does the optimal policy.

CC823 Room MAL 352 CONTRIBUTIONS IN FINANCIAL MARKETS Chair: Giacomo Bormetti

C0732: **The accuracy of trade classification algorithms: Evidence from the Budapest, Prague and Warsaw stock exchanges**
Presenter: **Sabina Nowak**, University of Gdansk, Poland

The aim is to evaluate and compare the accuracy of the transaction classification rules, including tick, reverse tick, quote, Lee and Ready and Ellis, Michaely and O'Hara rules, on the three Central and Eastern European stock exchanges: the Warsaw Stock Exchange, the Prague Stock Exchange and the Budapest Stock Exchange. The transaction data of 50 companies, the constituents of the blue chip indices, namely WIG20, PX and BUX, from October 2018 to June 2019, are employed. Tick data is collected from the Refinitiv (previously Thomson Reuters) database. The problem of the classification of transactions initiated by sellers and buyers as well as its accuracy is widely addressed in a current literature on the financial market microstructure. Most of the research concerns the US financial markets; research on other markets is scarce and usually includes a single country, as Australia, Brazil, Germany, Taiwan or Turkey. There is a lack of similar research devoted to the European developing markets. The aim is to fill this cognitive gap with reference to the selected markets of the Central and Eastern Europe.

C1664: **Information dissemination across high-latency cryptocurrency markets**
Presenter: **Thomas Dimpfl**, University of Tubingen, Germany
Co-authors: Dirk Baur

The time between order submission and confirmation is crucial for high frequency traders as they risk slippage when latency is too high. We hypothesize that high latency in cryptocurrency markets implies zero correlations of returns across exchanges at high frequencies as information is transmitted very slowly between them. To evaluate this conjecture, we measure the correlation of returns across exchanges at increasing sampling frequencies. Since all exchanges trade the same asset, correlations must eventually approach one. However, at high frequencies, correlations are close to zero. The strength of the correlation increases at a 10-minute frequency which is close to the median confirmation time (MCT) of the Bitcoin blockchain. To support the correlation findings, we use symbolic transfer entropy and find that information transmission is highest at frequencies between 5 and 15 minutes. In a second step, we explain the speed at which information is incorporated into prices using MCT and number of transactions (NoT) on the blockchain. We find that MCT is in general associated with slower information processing while the effect of NoT is ambiguous: for Bitcoin, more traffic seems to slow down the information processing while the reverse holds for the other cryptocurrencies. The findings highlight the link between cryptocurrency trading and the underlying blockchain technology and identify important differences of cryptocurrency trading and stock trading.

C0802: **A time-frequency analysis of financial market contagion in Europe**
Presenter: **Luis Aguiar-Conraria**, Universidade do Minho, Portugal
Co-authors: Maria Joana Soares, Mustapha Ojo

The increasing interconnectedness of the global economy and the rapid integration of global financial markets promote global economic growth, increase sheer volume and velocity of international financial transactions, and improve capital flow to many countries. On the flip side, it poses challenges to global economic and financial architectures. For instance, global financial markets have witnessed many financial and currency crises in the last four decades. One central feature of these crises is the snowballing effect from one market or geographical location to another. This largely undesired domino effect of financial market crises is generally termed contagion; there appears to be no consensus on its definition or measurements. Some prominent definitions, among several others, include a significant increase in the conditional probability of a crisis in one country relative to another country; focuses on the volatility spillover of asset prices from the crisis country to other countries; occurs when fundamentals cannot explain cross-country co-movements of asset prices. Financial contagion in the debt markets is investigated during various crisis-ridden periods in Eurozone using a wavelet approach and proposes a definition compatible with the time-frequency framework of wavelet analysis.

C0947: **Financial market switching-points and economic anomalies: Evidence from S&P100**
Presenter: **Takayuki Mizuno**, National Institute of Informatics, Japan
Co-authors: Yuan Yuan

The aim is to clarify the statistical relationship between the switching points of financial markets and various economic indicators by using machine learning. Using an extensive set of macroeconomic variables, LASSO regression is applied to select the important variables which have important impact on switching-points, then the switching-points over the period 2002-2016 of S&P100 is predicted by using the selected variables. The results show that selected variables have good out-of-sample predictive power. Moreover, ridge regression is also applied to analyze the switching-points. The contribution is to both anomaly detection of financial markets, and the application high dimensional data and machine learning to financial

market.

CG479 Room MAL 351 CONTRIBUTIONS IN COMMODITIES FINANCE Chair: Neil Kellard

C1408: **Commodity financialization and the role of speculators**
Presenter: **Xin Jin**, University of Aberdeen, United Kingdom

The aim is to investigate empirically the role of speculators in commodity financialization, which is an influx of capital into the commodity futures market and a rise of financial investors who are not interested in the physical commodities. Building a micro-foundation equilibrium model for commodity futures markets, it is shown that the speculative positions could affect both the commodity spot and futures prices. In particular the number of speculative traders drives the magnitude of the speculators effect on the prices. This hypothesis is tested using disaggregated Commitments of Traders report from Commodity Futures Trading Commission.

C1865: **A noncausal analysis of the role of expectations in forecasting energy prices**
Presenter: **Arthur Thomas**, IFP Energies nouvelles-University of Nantes, France

The aim is to provide evidence of the pivotal role of expectations to forecast energy prices through the estimation of a bivariate Bayesian non-causal VAR model which includes real-time energy prices and the associated convenience yield. Convenience yield in oil and gas markets is an implied return on holding inventories and thus proxies for the interest market participants have for the commodity. The approach significantly reduces the number of estimated parameters compared to the reference models in the literature. We show that real-time forecasts of real energy prices can be much more accurate than the no-change forecast and notably more accurate than real-time forecasts generated by existing models relying on Bayesian VAR. These results are in striking contrast with those published in the literature on asset prices.

C1911: **Commodity futures: Does the traded volume influence research interest?**
Presenter: **Isabel Catalina Figuerola-Ferretti Garrigues**, Universidad Pontificia Comillas, Spain
Co-authors: Ioannis Paraskevopoulos, M Teresa Corzo, Karin Martin-Bujack

The aim is to establish the process of building up knowledge from commodity markets using market observables from commodity futures trading activity. We analyze the way in which knowledge feeds back into market prices via the channels of different regulation policies. Employing a textual analysis, we build our proxies of knowledge using the total number of publications in commodities and oil specific markets and run a cointegration test for the 2000-2018 period, against two measures of trading activity. Academics observe prices and volumes traded and respond with more research. For oil, the number of papers published are substantially higher than the number of papers found under the commodity text search. We also find strong cointegration between research activity and the dollar value of futures trading volumes in the oil market. We split the sample into two subperiods to take account of the effects of the Global Financial Crises. Knowledge explains market activity under our crude oil market measures during the post-crises period.

C1635: **Capital mobility in commodity-exporting economies**
Presenter: **Konstantin Rybak**, RANEPA, Russia
Co-authors: Andrey Zubarev, Andrey Polbin

Capital mobility is studied in commodity-exporting economies. These countries substantially depend on world commodity prices and have rather high level of savings on average, so it is naturally to assume that they demonstrate special patterns of capital mobility. The main hypothesis is that constraints on capital mobility in this group of countries depend upon the level of savings compared to the level of investments. In particular, with high savings that follow higher world demand and higher commodity prices, financing country's desirable level of investment is not a big deal. At the same time, in the case of negative terms of trade shocks these commodity-exporting economies may experience lower savings and higher country risk-premium. This may lead to restrictions on borrowing capital in the global market, resulting in a high correlation between investments and savings. The results of threshold regressions speak in favour of our hypothesis.

CG692 Room MAL 353 CONTRIBUTIONS IN SENTIMENT ANALYSIS Chair: David Ardia

C1729: **Fine-grained, aspect-based semantic sentiment analysis on news for economic forecasting and nowcasting**
Presenter: **Sergio Consoli**, European Commission, Joint Research Centre, Italy
Co-authors: Sebastiano Manzan, Luca Barbaglia

News are a promising nowcasting and forecasting tool since they describe current economic events and the expectations of economic agents about the future. In particular understanding the sentiment embedded in current economic news may provide additional signals to improve forecasts of economic models and produce more accurate predictions. Recent work in economics on the application of sentiment analysis from social media and news generally suffers from: (i) a limited scope of historical financial news available; (ii) analysis of short texts only (e.g. usually tweets or news headlines); and (iii) use of basic text analysis techniques. We provide an overview on the development of a fine-grained, aspect-based sentiment analysis approach. The method is based on advanced natural language processing and is able to recognize the exact entity to which the sentiment aspect is expressed within entire, long text articles. The approach is unsupervised since it relies on external lexical resources to associate a polarity score to each concept. We describe our novel method and report some preliminary findings on historical economic news which consider a long time period of 25 years. Our analysis provides evidence on the usefulness of considering news sentiment as an additional instrument in the forecasting toolset of macroeconomists.

C1735: **Environmental sentiment in financial market: A global warning**
Presenter: **Marie Bessec**, University Paris Dauphine, France
Co-authors: Julien Fouquau

There is growing literature using media text to capture investors' sentiment and investigating its influence on financial markets. We use textual analysis to measure the growing concern about climatic issues and to assess its impact on stock prices in the United States. Using a dataset of 71,785 articles published in the Wall Street Journal from 2010 to 2019, we create several environmental scores capturing the media coverage for environmental issues and we investigate their influence on 494 S&P500 constituents in a Fama-French five factor model. We do find a significant impact of the environmental sentiment on the stock returns of nearly 25% of the firms. The response varies across the different sectors. As expected, the effect is negative in energy and materials, in particular in chemicals and metals. A positive impact is found in retail estate and utilities. These results are robust to the use of alternative lexicons, various term weighting and sample periods.

C0970: **How to gauge investor behavior: A comparison of online investor sentiment measures**
Presenter: **Simon Behrendt**, Zeppelin University, Germany
Co-authors: Daniele Ballinari

Given the increasing interest in and the growing number of publicly available methods to estimate investor sentiment from social media platforms, researchers and practitioners alike are facing one important question - which is best to gauge investor sentiment? We compare the performance of daily investor sentiment measures estimated from Twitter and StockTwits short messages by publicly available dictionary and neural network based methods for a sample of 360 stocks over a seven years time period. To determine their relevance for financial applications, these investor sentiment measures are compared by (i) their effect on the cross-section of returns and (ii) their ability to forecast abnormal portfolio returns and

trading volume. We provide a clear ranking of the considered online investor sentiment measures, elaborate on the reasons for the differences in performance across measures and add a note on the well-known reversal effect.

C1847: **Emotions in macroeconomic news and their impact on the European bond market**

Presenter: **Luca Tiozzo Pezzoli**, JRC European Commission, Italy

Co-authors: Sergio Consoli, Elisa Tosetti

The aim is to study how emotions extracted from macroeconomic news can be used to explain and forecast future behaviour of sovereign bond yield spreads in Italy and Spain in the period from 2 March 2015 to 31 December 2018. We use a big, open-source, news-level database known as Global Database of Events, Language and Tone to construct a set of variables capturing daily variations in the emotional content of news about national and foreign economic events and use them as a proxy for market investor's expectations and behaviour. We focus on a set of negative emotions, conveying varying intensities of fear and anxiety, and linked to the different affective states of investors. We find that negative emotions such as anxiety and panic extracted from news are good predictors for the increase in sovereign bond yield spread for the two countries. Relatively stronger emotions, such as panic, reveal useful information for forecasting one-day ahead and five-day ahead changes in spread. Finally, we find that negative emotions generated by the Italian political turmoil in the period June-December 2018 propagated to the Spanish news, affecting the Spanish sovereign bond yield spread.

Monday 16.12.2019 08:40 - 10:20 Parallel Session M – CFE-CMStatistics

EO136 Room CLO B01 RECENT DEVELOPMENTS IN FUNCTIONAL DATA ANALYSIS Chair: Sara Lopez Pintado

E1145: **Consistent bootstrap procedures for testing on the coefficients of a Hilbert-valued regression model**
Presenter: **Gil Gonzalez-Rodriguez**, University of Oviedo, Spain
Co-authors: Ana Colubi

A convenient mathematical framework to deal with general random elements is a separable Hilbert space. This is the case, in many situations, of functional-valued random variables considered in functional data analysis. In this framework non-parametric techniques are specially useful due to the scarcity of parametric distributions. Particularly, bootstrap has proved to be highly valuable to develop inferential procedures. A quite general methodology to theoretically prove the consistency of many bootstrap approaches for random elements taking values in separable Hilbert spaces has been recently introduced. Within this context, a linear regression model with real-valued explanatory variables and Hilbert-valued coefficients and response will be considered. Several bootstrap procedures will be proposed to test about general combinations of the coefficients of this model. Its consistency will be proved by applying the general methodology.

E1275: **Functional registration of walking strides in high-density accelerometry data for estimation of gait asymmetry**
Presenter: **Marta Karas**, Johns Hopkins Bloomberg School of Public Health, United States
Co-authors: Ciprian Crainiceanu, Jacek Urbanek, Amy Bastian, Ryan Roemmich

Ability to characterize human gait pattern has significant potential in health research and can help guide clinical decision making. For example, in stroke survivors, quantification of asymmetry in walking strides during and after an in-lab intervention is of particular interest. To address this problem, we use high-density, high-throughput wearable accelerometry data and propose a stride pattern registration framework. We use a two-parameter family of time-warping functions to estimate clinically relevant stride characteristics, including duration and asymmetry. To demonstrate the approach, we collect accelerometry data on a healthy adult walking on a split-belt treadmill under different conditions mimicking step-to-step asymmetry. To analyze the data, we first use ADaptive Empirical Pattern Transformation (ADEPT) - a fast and scalable method for strides segmentation. We then employ a parametrized stride pattern framework to further characterize segmented strides. We conclude that the parametrized adaptive pattern matching appears to be a promising approach for estimation of step asymmetry.

E1456: **Statistical analysis of longitudinal data on Riemannian manifolds**
Presenter: **Xiongtao Dai**, Iowa State University, United States
Co-authors: Zhenhua Lin, Hans-Georg Mueller

A manifold version of the principal analysis by conditional expectation (PACE) is proposed to represent sparsely observed longitudinal data that take values on a nonlinear Riemannian manifold. Typical examples of such manifold-valued data include longitudinal compositional data, as well as longitudinal shape trajectories located on a hypersphere. Compared to standard functional principal component analysis that is geared to Euclidean geometry, the proposed approach leads to improved trajectory recovery on nonlinear manifolds in simulations. As an illustration, we apply the proposed method on longitudinal emotional well-being data for unemployed workers. An R implementation of our method is available on GitHub.

E1255: **Asymptotics and regularization in spherical functional autoregressive models**
Presenter: **Alessia Caponera**, Sapienza University of Rome, Italy

A class of space-time random fields, which are functional autoregressive processes taking values in the space of square integrable functions on the sphere, is presented. We discuss the estimation of the corresponding autoregressive kernels. In particular, we investigate asymptotic properties of a form of least squares regression and LASSO-type estimators.

EO142 Room Bloomsbury RECENT ADVANCES IN SURVIVAL ANALYSIS Chair: Chiung-Yu Huang

E0322: **Analysis of restricted mean survival time for length-biased data**
Presenter: **Chi Hyun Lee**, University of Massachusetts Amherst, United States
Co-authors: Jing Ning, Yu Shen

In clinical studies with time-to-event outcomes, the restricted mean survival time (RMST) has attracted substantial attention as a summary measurement for its straightforward clinical interpretation. When the data are subject to length-biased sampling, which is frequently encountered in observational cohort studies, existing methods to estimate the RMST are not applicable. We consider nonparametric and semiparametric regression methods to estimate the RMST under the setting of length-biased sampling. To assess the covariate effects on the RMST, a semiparametric regression model that directly relates the covariates and the RMST is assumed. Based on the model, we develop unbiased estimating equations to obtain consistent estimators of covariate effects by properly adjusting for informative censoring and length bias. Stochastic process theories are used to establish the asymptotic properties of the proposed estimators. We investigate the finite sample performance through simulations and illustrate the methods by analyzing a prevalent cohort study of dementia in Canada.

E1000: **Accelerated failure time model based on nonparametric Gaussian scale mixtures**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South
Co-authors: Sangwook Kang

When some parametric error distributions, such as normal for the accelerated failure time model, are assumed, estimators typically suffer from misspecification problems. To relax this problem, we propose a nonparametric Gaussian scale mixture model to flexibly specify the error distribution. Unlike existing non- or semi-parametric estimation methods such as rank-based procedures, the proposed method enables us to use an explicit likelihood function while avoiding potential misspecification problems. We present this model with specific estimating algorithms and some numerical examples.

E1763: **Nonparametric maximum likelihood estimation of acceleratedfailure time models for competing risks**
Presenter: **Sangbum Choi**, Korea University, Korea, South

Competing risks are common in clinical cancer research, as patients are subject to multiple potential failure outcomes, such as death from the cancer itself or from complications arising from the disease. In the analysis of competing risks, several regression methods are available for the evaluation of the relationship between covariates and cause-specific failures, many of which are based on Cox's proportional hazards model. Although a great deal of research has been conducted on estimating competing risks, less attention has been devoted to linear regression modeling, which is often referred to as the accelerated failure time (AFT) model in survival literature. We propose maximum likelihood inference procedures based on the kernel-smoothing principle for the AFT model under competing risks scenarios. The estimator is shown to be consistent and asymptotically normal. The performance of the new inference procedures is assessed through simulation studies, where the proposed estimator and the estimator from a cause-specific model are also compared. Illustrations with data from non-Hodgkin lymphoma patients are provided.

E1943: **Induced smoothed methods in analysis of semiparametric quantile residual lifetimes models**
Presenter: **Sangwook Kang**, Yonsei University, Korea, South
Co-authors: Kyuhyun Kim

The focus is on statistical inference procedures for fitting a semiparametric quantile residual life (SQRL) model that models life expectancy.

Quantile residual lifetimes are essential summary measures in survival analysis along with a hazard function or survival function. Recent statistical inference procedures for fitting SQRL models have been estimating functions approaches that is nonsmooth in model parameters. Thus, optimizing objective functions or solving estimating equations could be very cumbersome. We propose to employ a computationally-efficient induced-smoothing procedure that smoothes nonsmooth estimating functions. Variance estimation can be done via efficient resampling procedures that uses the sandwich form of asymptotic variances. We establish the consistency and asymptotic normality of the proposed estimators. Finite sample properties are investigated via an extensive simulations studies. We illustrate our proposed methods with a real dataset.

EO138 Room G11 RECENT DEVELOPMENTS OF STATISTICAL METHODS FOR CAUSAL INFERENCE Chair: Zheng Zhang

E0527: **Regularized calibrated estimation and model-assisted inference for treatment effects with high-dimensional data**
Presenter: **Zhiqiang Tan**, Rutgers University, United States

Consider the problem of estimating average treatment effects in the framework of potential outcomes when a large number of covariates are used to adjust for possible confounding through outcome regression and propensity score models. We develop new methods and theory to obtain doubly robust point estimators for average treatment effects, which remain consistent if either the propensity score model or the outcome regression model is correctly specified. We also obtain model-assisted confidence intervals, which are valid when the propensity score model is correctly specified, but the outcome regression model may be misspecified. Our methods involve regularized calibrated estimators with Lasso penalties, but carefully chosen loss functions, for fitting propensity score and outcome regression models. We provide high-dimensional analysis to establish the desired properties of our methods under comparable sparsity conditions to previous results, which give valid confidence intervals when both the propensity score and outcome models are correctly specified. We present simulation studies and an empirical application which demonstrate advantages of the proposed methods compared with related methods based on regularized maximum likelihood estimation.

E0562: **Optimal causal inference in a high-dimensional discrete model**
Presenter: **Edward Kennedy**, Carnegie Mellon University, United States

Estimation of the average causal effect is considered in a setting where the covariates required for confounding adjustment are discrete but arbitrarily high-dimensional. To the best of our knowledge, this framework has yet to be studied. We provide non-asymptotic risk bounds for a standard plug-in estimator, showing that this estimator is only consistent in the regime where the dimension grows slower than the sample size. Then we go on to characterize the minimax lower bound. We also consider several variations on this setup: one where we weaken the classical positivity assumption so that the bound on propensity scores can become more extreme with sample size (in which the functional becomes nonsmooth and the minimax rates change substantially), one where we instead target a data-adaptive causal effect conditional on being in a high-probability category, and one where we consider a sparsity condition that limits the heterogeneity of conditional causal effects.

E1460: **Causal inference with confounders missing not at random**
Presenter: **Shu Yang**, North Carolina State University, United States
Co-authors: Linbo Wang, Peng Ding

It is important to draw causal inference from observational studies, which, however, becomes challenging if the confounders have missing values. Generally, causal effects are not identifiable if the confounders are missing not at random. We propose a novel framework to nonparametrically identify causal effects with confounders subject to outcome-independent missingness, that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. We then propose a nonparametric two-stage least squares estimator and a parametric estimator for causal effects.

E1544: **A simple and efficient estimation of average treatment effects in the presence of unmeasured confounders**
Presenter: **Zheng Zhang**, Renmin University of China, China

A critical condition in the treatment evaluation literature is that, conditional on all confounders, participation decision is independent of potential outcomes. If some of those confounders are not observable, then conditional on the observable confounders, participation decision is no longer independent of potential outcomes and consequently the average treatment effect (ATE) is not identified without further assumption. Indeed, the literature establishes that ATE is not identified even if standard instrumental variables are available. Two proposals are suggested. The first one assumes that the unmeasured confounders are not interacted with the treatment in potential outcomes or with the instrument in participation decision. Under this additional restriction, ATE is identified. The other assumes that the complier's treatment status is monotone in instrument. Under this additional restriction, the local average treatment effect (LATE) for the compliers is identified. A simple and efficient estimation of ATE and LATE is proposed which does not estimate the influence function parametrically, thereby is more robust than the existing methods.

EO632 Room G21A MODELLING DEPENDENCE THROUGH GRAPHICAL MODELS Chair: Ghislaine Gayraud

E0411: **Labelled dynamic Bayesian network: Framework and structure learning for application to ecological network**
Presenter: **Etienne Auclair**, UTeam, France

A "labelled dynamic Bayesian network" is a particular case of dynamic Bayesian network where the probabilities of the random variables are not described by conditional probabilities tables, but by a small set of parameters shared by every variable of the network. The interest of this model is the fact that the number of parameters of the model is fixed, and does not depend on the size and the complexity of the network. This model will be described with an algorithm developed to learn the structure of a labelled dynamic Bayesian network, illustrated on an ecological problem: learning the structure of an ecological network using dynamic data of presence/absence of the species.

E0412: **Learning structures of Bayesian networks with cyclic structures**
Presenter: **Witold Wiecek**, Certara UK Ltd, United Kingdom
Co-authors: Frederic Bois, Ghislaine Gayraud

Bayesian networks are a popular approach to modelling networks. Networks in BNs must be acyclic while in many applications they include cycles. Dynamic BNs can be used but they require time series data. We present an alternative model that embeds cyclic structures within acyclic BNs, allowing us to still use the factorization property of BNs and informative priors on network structure. We present an implementation in the linear Gaussian case, where cyclic structures are treated as multivariate nodes. We use a Markov Chain Monte Carlo algorithm for inference, allowing us to work with the whole posterior distribution on the space of graphs. The algorithm implemented as a part of graph_sampler, open-source software for modelling networks.

E0417: **Application of dynamic Bayesian network approaches for quantitative adverse outcome pathway modelling**
Presenter: **Wang Gao**, University of Technology of Compiegne, France
Co-authors: Ghislaine Gayraud, Frederic Bois

In toxicology, an Adverse Outcome Pathway (AOP) is a conceptual framework that qualitatively describes the existing knowledge on the links between the two anchor points: Molecular Initiating Event (MIE) and Adverse Outcome (AO) at a level of biological organization relevant for risk assessment. The transformation of an AOP to its quantitative version, qAOP allows building a powerful risk assessment tool, thanks to its ability to quantitatively predict the AO. Given that an AOP is by definition a directed chain describing toxicological causality, we propose new methods based on dynamic Bayesian networks for qAOP modelling. The linear and non-linear qAOP models based on different assumptions (stochastic transition without observational error or deterministic transition with Gaussian observational errors) will be introduced in the presentation. We

will demonstrate and compare the numerical results of our models applied to simulated data and real data from the toxicological studies of chronic kidney disease and Parkinsonian motor deficits.

E0439: **Non-homogeneous dynamic Bayesian networks with Bayesian regularization for gene regulatory network inference**
Presenter: **Sophie Lebre**, IMAG, France
Co-authors: Frank Dondelinger, Dirk Husmeier

The proper functioning of any living cell relies on complex networks of gene regulation. These regulatory interactions are not static, but respond to changes in the environment and evolve during the life cycle of an organism. A challenging objective in computational systems biology is to infer these time-varying gene regulatory networks from typically short time series of transcriptional profiles. While homogeneous models, like conventional dynamic Bayesian networks, lack the flexibility to succeed in this task, fully flexible models suffer from inflated inference uncertainty due to the limited amount of available data. We explore here a semi-flexible model based on a piecewise homogeneous dynamic Bayesian network regularized by gene-specific inter-segment information sharing. We consider different choices of prior distribution and information coupling, and evaluate their performance on synthetic data and gene expression time series.

EO616 Room G3 MULTIVARIATE EXTREMES AND CAUSALITY Chair: Johanna Neslehova

E0621: **Causal discovery for heteroscedastic financial series**
Presenter: **Valerie Chavez-Demoulin**, University of Lausanne, Switzerland

Inferring causality between financial assets is a common and fundamental subject in finance. The widely used Granger causality allows us to determine whether one time series is useful in forecasting another. Under Granger causality, the cause happens prior to its effect. We propose a method to understand intrinsic causal mechanisms between series, unconditionally on time. Dealing with heteroscedastic financial data, we investigate causal relations not only in mean but from the perspective of location, scale and shape parameters of the underlying distribution. The method is called causal heteroscedastic model(CHM) and it admits non-linear and non-Gaussian causal multiplicative noise models. We show its performance based on an extensive simulation study and apply it to the extreme returns of financial times series to discover intrinsic causal mechanisms under regimes of crash.

E0645: **Causal mechanism of extreme river discharges in the upper Danube basin network**
Presenter: **Linda Mhalla**, HEC Montreal, Canada
Co-authors: Valerie Chavez-Demoulin, Debbie Dupuis

Extreme hydrological events in the Danube river basin have tragic consequences for human populations, living aquatic organisms, and the economic activity. One often characterizes the joint structure of the extreme events using the theory of multivariate and spatial extremes and its asymptotically justified models. There is interest however in cascading extreme events and whether one event causes another. We argue that an improved understanding of the mechanism underlying severe events is achieved by combining extreme value modelling and causal discovery. We construct a causal inference method relying on the notion of Kolmogorov complexity of extreme conditional quantiles. Tail quantities are derived using multivariate extreme value models and causal-induced asymmetries in the data are explored through the minimum description length principle. Owing to the developed methodology, we uncover causal relations between summer extreme river discharges in the upper Danube basin and find significant causal links between the Danube and its Alpine tributary Lech.

E0743: **Nonasymptotic analysis of the angular measure for extremes, application to classification**
Presenter: **Anne Sabourin**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Hamid Jalalzai, Stephan Clemencon

In multivariate extreme value theory, the angular measure characterizes the dependence structure of multivariate heavy-tailed variables. In the case where the components have different tail indices, standardization using the rank-transformation (empirical distribution function) is a common practice. We propose a modification of the classical empirical estimator based on the rank-transformed sample, based on intermediate data, ie. upon data which norm rank among the largest of the observed sample, but not among the very largest. In other word we discard the very largest data. We provide a nonasymptotic bound for the uniform deviations of the empirical angular measure evaluated on rectangles of the unit sphere. Our bound scales as the squared root of the number of observations used for inference. This nonasymptotic study is, to the best of our knowledge, the first of its kind in this domain. As an application, we provide finite sample guarantees for classification in extreme regions and anomaly detection via minimum-volume sets estimation on the sphere.

E1362: **Modelling multivariate observations from the bulk of the data to the extremes**
Presenter: **Anne-Laure Fougeres**, Universite Claude Bernard - Lyon, France
Co-authors: Johanna Neslehova, Simon Chatelain

The aim is to discuss how to model multivariate phenomena such that not only the extremes are of particular interest, but also the medium regime. An application to rainfall data over France will be used to illustrate the procedures introduced.

EO174 Room G4 ROBUST STATISTICS Chair: Tim Verdonck

E0752: **Clustering dynamic panels of income data with robust interactive fixed effects**
Presenter: **Kris Boudt**, Vrije Universiteit Brussel, Belgium
Co-authors: Ewoud Heyndels

The framework of interactive fixed effects is extended to make it robust against the presence of outliers. In each iteration the coefficients of the observable variables are estimated with robust regressions and the unobserved factors are extracted by calculating a robust covariance matrix and using this to determine robust eigenvectors. We apply the method to cluster income time series of Belgian independent entrepreneurs and to determine the unobserved factors.

E1356: **Robust discriminant analysis for high dimensions**
Presenter: **Valentin Todorov**, UNIDO, Austria
Co-authors: Peter Filzmoser

The classical discriminant methods (LDA and QDA) could suffer from the singularity problem in cases of high-dimensional small sample size data, which limits their practical application. A number of regularization techniques with the purpose to stabilize the classifier and to achieve an improved classification performance have been developed and there exist several studies comparing various regularization techniques trying to facilitate the choice of a method. However, none of these methods takes into consideration the possible presence of outliers in the training data set which can strongly influence the obtained classification rules and make the results unreliable. On the other hand, the high breakdown point versions of discriminant analysis (with one exception) proposed in the literature do not work or are not reliable in high dimensions. The method we propose relies on the recently introduced regularized versions of the minimum covariance determinant (MCD) estimator - RMCD and MRCD - and combines high robustness to outliers, the possibility to be computed for high dimensions and readily available software in R. Simulated and real data examples show that the proposed method performs better than, or at least as well as, the existing methods in a wide range of settings.

E1410: **Fast robust correlation for high dimensional data**
Presenter: **Jakob Raymaekers**, KULeuven, Belgium

Co-authors: Peter Rousseeuw

The product moment covariance is a cornerstone of multivariate data analysis, from which one can derive correlations, principal components, Mahalanobis distances and many other results. Unfortunately the product moment covariance and the corresponding Pearson correlation are very susceptible to outliers (anomalies) in the data. Several robust measures of covariance have been developed, but few are suitable for the ultrahigh dimensional data that are becoming more prevalent nowadays. For that one needs methods whose computation scales well with the dimension, are guaranteed to yield a positive semidefinite covariance matrix, and are sufficiently robust to outliers as well as sufficiently accurate in the statistical sense of low variability. We construct such methods using data transformation. The resulting approach is simple, fast and widely applicable. We study its robustness by deriving influence functions and breakdown values, and computing the mean squared error on contaminated data. Using these results we select a method that performs well overall, which we call wrapping and which is available in the R package cellWise. Wrapping allows a very substantial speedup of the DetectDeviatingCells technique for flagging cellwise outliers, which is applied to genomic data with 12,000 variables. Wrapping is able to deal with even higher dimensional data, which is illustrated on color video data with 920,000 dimensions.

E1615: **M-type penalized splines with auxiliary scale estimation**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst

Penalized spline smoothing is a popular and flexible method of obtaining estimates in nonparametric regression but the classical least-squares criterion is highly susceptible to model deviations and atypical observations. Penalized spline estimation with a resistant loss function is a natural remedy, yet to this day the asymptotic properties of M-type penalized spline estimators have not been studied. We show that M-type penalized spline estimators achieve the same rates of convergence as their least-squares counterparts, even with auxiliary scale estimation. We further find theoretical justification for the use of a small number of knots relative to the sample size. We illustrate the benefits of M-type penalized splines in a Monte-Carlo study and two real-data examples, which contain atypical observations.

EO705 Room G5 RECENT ADVANCES IN DIMENSION REDUCTION Chair: Eliana Christou

E1009: **Stationary subspace analysis: A statistical perspective**
Presenter: **Klaus Nordhausen**, Vienna University of Technology, Austria
Co-authors: Lea Flumian, Markus Matilainen

Multivariate times series occur in many application areas and are challenging to model. A common approach is therefore to assume that the observed time series can be decomposed into latent components with different exploitable properties. In some of these models especially nonstationary components are of interest, and thus the nonstationary subspace should be separated from the stationary subspace which is often referred to as stationary subspace analysis (SSA). Different methods are considered for this purpose and a test suggested to make inference about the dimensions of the subspaces.

E0356: **Moment kernel for estimating central mean subspace and central subspace**
Presenter: **Xiangrong Yin**, University of Kentucky, United States
Co-authors: Weihang Ren, Dennis Cook

The T-central subspace allows one to perform sufficient dimension reduction for any statistical functional of interest. We propose a general estimator using (third) moment kernel to estimate the T-central subspace. We particularly focus on central mean subspace via the regression mean function, and central subspace via Fourier transform or slicing. Theoretical results are established and simulation studies show the advantages of our proposed methods.

E0385: **Adaptively weighted large margin classifiers for sufficient dimension reduction**
Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom

Adaptively weighted large margin classifiers are combined with Support Vector Machine (SVM)-based dimension reduction methods to create dimension reduction methods robust to the presence of extreme outliers. We discuss estimation and asymptotic properties of the algorithm. The good performance of the new algorithm is demonstrated through simulations and real data analysis.

E0389: **Nonparametric graphical models for high-dimensional functional data**
Presenter: **Eftychia Solea**, Ruhr Universitat Bochum, Germany
Co-authors: Holger Dette

We consider the problem of constructing nonparametric undirected graphical models for multivariate functional data. Most existing approaches on graphical models assume either the Gaussian distribution on the vertices or linear conditional means. The presented approach provides a more flexible model which relaxes the linearity assumption by replacing it by an arbitrary additive form. The utilisation of the functional principal components offers an estimation strategy that uses a group lasso penalty to estimate the relevant edges of the graph. We establish the model selection consistency for the resulting estimator, while allowing both the number of predictors and the number of functional principal components to diverge to infinity with the sample size. We investigate the empirical performance of our method through simulation studies and a real data application.

EO717 Room Gordon EFFICIENT AND OPTIMAL DESIGN OF EXPERIMENTS Chair: Heiko Grossmann

E0722: **New approaches for experiments with mixtures**
Presenter: **Stefanie Biedermann**, University of Southampton, United Kingdom
Co-authors: Steven Gilmour, Rana Khashab

Experiments involving mixtures are conducted in a variety of areas, for example in food processing or in chemical research. The experimental region is constrained naturally, as the proportions of all ingredients have to sum to one. Additional constraints may arise when there are bounds on the proportions, for example a cake must contain a minimum percentage of flour to have the right texture and flavour. We propose a new - parsimonious but flexible - class of non-linear models, based on fractional polynomials, to fit the data from constrained mixture experiments. We will motivate this modelling approach, and will use a number of historical data sets to compare these models with various other models suggested in the literature. We will then present some optimal designs for these models, and will further discuss some general issues related to designing experiments for mixtures.

E1072: **On greedy heuristics for computing D-efficient saturated subsets**
Presenter: **Radoslav Harman**, Comenius University in Bratislava, Slovakia
Co-authors: Samuel Rosa

D-efficient saturated subsets are natural initial solutions of various algorithms applied in statistics and computational geometry. We propose two greedy heuristics for the construction of D-efficient saturated subsets: an improvement of a previous method in the context of D-optimal experimental designs, and a modification of another method for the initiation of the minimum-volume enclosing ellipsoid algorithms. We provide mathematical insights into the two methods, and compare them to the commonly used random and regularized heuristics. We also demonstrate the speed of the state-of-the-art algorithms for computing D-optimal approximate designs of experiments initiated by the proposed methods.

E1207: **Bayesian design of physical experiments for nonlinear and computational models**
Presenter: **Tim Waite**, University of Manchester, United Kingdom
Co-authors: David Woods, Yiolanda Englezou
The purpose is to discuss Bayesian decision-theoretic optimal design of physical experiments for parameter estimation in nonlinear models, including models that incorporate an expensive computer simulator. In the Bayesian approach, one key challenge is the presence of analytically intractable nested integrals in the expected utility (e.g. expected Shannon information gain) of any proposed design. We propose new Monte Carlo approaches for approximate numerical integration of the expected utility that give reduced bias and computational expense compared to several existing methods. Another challenge is that, when the model incorporates an expensive computer simulator, it is prohibitively costly to use an expected utility estimate that relies on direct evaluations of the simulator. Hence in order to perform design optimization for the physical experiment, and also to conduct subsequent inference, one must use a computationally cheap surrogate model in place of the simulator. We accomplish this using a Gaussian process emulator built with pre-existing training data from a computer experiment, thereby extending the analysis framework from the calibration literature to the design problem. The proposed fully Bayesian framework enables uncertainty about the simulator output at untested input combinations to be incorporated when designing the physical experiment.

E1551: **Subdata selection methods**
Presenter: **John Stufken**, University of North Carolina at Greensboro, United States
The size of big data can cause challenges for even the simplest explorations of the data. Such challenges can, for example, be related to storage of the data or to computations of even the simplest statistics. One method to deal with the challenges is based on selecting a much smaller subdata set from the original full data set. Exploration or analysis would proceed with the subdata. Such subdata set can be selected through a sampling strategy or through a deterministic method that attempts to optimize a specified criterion. Whatever method of subdata selection is used, it is important that it is computationally feasible and efficient. It is also important that inferences or predictions based on the subdata are comparable to those that would have been obtained by using the full data. Ideally, this is true with as few assumptions as possible about the full data. After a brief discussion of different subdata selection methods, we will focus on comparison of the methods, their strengths and weaknesses, and possible extensions.

EO086 Room MAL G13 LEARNING AND INFERENCE METHODOLOGIES FOR STOCHASTIC PROCESSES Chair: Hiroki Masuda

E0338: **Estimation for partially observed epidemic dynamics with measurement errors**
Presenter: **Maud Delattre**, AgroParisTech, France
Co-authors: Catherine Laredo, Romain Narci, Elisabeta Vergu
Estimating the parameters governing epidemic dynamics, such as the transmission rate, from available data is a major issue in order to provide reliable predictions of these dynamics and of the impact of control strategies. In this context, several difficulties occur: all the components of the system dynamics are not observed, and data are available at discrete times with measurement errors. Diffusion processes with small diffusion coefficient are a convenient set-up for modelling epidemics, the small diffusion coefficient being related to the population size. In practical applications on epidemic dynamics, it often occurs that some coordinates of the diffusion are not observed and, when observed, measurement errors are systematically present. We are then concerned with the estimation of the parameters when the diffusion process is discretely observed with noise on a finite time interval, and when some components cannot be observed. We propose a procedure derived from Kalman filtering approaches to compute estimates of the parameters based on approximate likelihoods. Our approach is original because it combines the framework of diffusions with small diffusion coefficient with approximate likelihood methods and Kalman filtering, the latter being little exploited for the inference of epidemic dynamics partially observed and with errors.

E0670: **De-biased graphical Lasso for high-frequency data**
Presenter: **Yuta Koike**, University of Tokyo, Japan
A new statistical inference theory is developed for the precision matrix of high-frequency data in a high-dimensional setting. The focus is not only on point estimation, but also on interval estimation and hypothesis testing for entries of the precision matrix. To accomplish this purpose, we establish an abstract asymptotic theory for the weighted graphical Lasso and its de-biased version without specifying the form of the initial covariance estimator. We also extend the scope of the theory to the case that a known factor structure is present in the data. The developed theory is applied to the concrete situation where we can use the realized covariance matrix as the initial covariance estimator, and we obtain a feasible asymptotic distribution theory to construct (simultaneous) confidence intervals and (multiple) testing procedures for entries of the precision matrix.

E0854: **Parametric inference for a parabolic SPDE from discrete observations**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Co-authors: Yusuke Kaino
The focus is on the estimation problem of unknown parameters for a parabolic linear second order stochastic partial differential equation (SPDE) based on high-frequency data which are observed in time and space. Previously, the parabolic linear second order SPDE model based on high-frequency data observed on a fixed region has been studied. The asymptotic properties of least squares estimators has been proved for both the normalized volatility parameter and the curvature parameter. We propose adaptive maximum likelihood (ML) type estimators of the coefficient parameters including the volatility parameter of the parabolic linear second order SPDE model by using thinned data obtained from high-frequency data. It is also shown that the adaptive ML type estimators have asymptotic normality under some regularity conditions. Furthermore, in order to verify asymptotic performance of the adaptive ML type estimators of the coefficient parameters of the parabolic linear second order SPDE model based on high-frequency data, some examples and simulation results of the adaptive ML type estimators are given.

E1006: **Finite mixture approximation of CARMA model**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
The aim is to show how to approximate the transition density of a CARMA(p,q) model driven by a time changed Brownian motion based on the Laguerre polynomial. We apply this result in two situations. Firstly we derive an analytical formula for option prices when the log price follows a CARMA model. We also propose an estimation procedure based on the approximated likelihood density.

EO737 Room MAL G14 BAYESIAN INFERENCE Chair: Jeff Hart

E0325: **The role of sparsity and misspecification in Bayesian model selection**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
Co-authors: Francisco Javier Rubio
The state-of-the-art in Bayesian model selection is to induce sparsity to ensure that, asymptotically, one is able to select the optimal model with probability one. Sparsity can be induced either via placing high prior probabilities on small models, setting large prior dispersion (diffuse priors) or using non-local priors. We present recent theoretical and empirical results showing potential adverse effects of using overly sparse priors. We also discuss how these issues are compounded with model misspecification, which we illustrate typically results in a loss of power to detect truly relevant signals. As part of our theoretical results we provide simple rates to help understand how fast one can recover the desired model. We also show a form of asymptotically valid uncertainty quantification for the selected model that is also valuable for L0-penalized regression, the effects

of failing to record relevant variables and potential issues with censored data in survival frameworks. We illustrate the practical relevance of these results via empirical studies.

E0920: **Cross-validation Bayes factors for the nonparametric two-sample test**
Presenter: **Jeff Hart**, Texas AM University, United States
Co-authors: Naveed Merchant, Taeryon Choi

Given independent random samples from densities f and g, a fundamental problem is testing equality of f and g. We define Bayes factors that utilize data splitting to test this hypothesis. Two models are considered: one, M_1, that assumes the densities are the same, and a second, M_2, that allows f and g to be different. Each data set is split into two parts, training and validation. Three kernel density estimates (KDEs) are computed from the training data, and the models M_1 and M_2 are defined in terms of these kdes. A marginal likelihood for each model is then computed from the validation data, and the Bayes factor is the ratio of the two marginal likelihoods. The relative simplicity of this method in comparison to existing nonparametric Bayes procedures is emphasized. Only three parameters are involved in the proposed method, these being the bandwidths of the three KDEs. Appropriate priors for the bandwidths are proposed, and the importance of choosing a good kernel for the KDEs is discussed. In particular, relatively heavy-tailed kernels should be used to guarantee good performance of the Bayes factors in a variety of settings.

E0953: **Bayesian spatial homogeneity pursuit methods**
Presenter: **Guanyu Hu**, University of Connecticut, United States

The Bayesian spatial homogeneity pursuit methods will be discussed. To capture the spatial homogeneity, we develop a Markov random fields constraint mixture of finite mixture prior. An efficient Markov chain Monte Carlo (MCMC) algorithm is designed to estimate parameters and their uncertainty measures simultaneously. Extensive simulations are conducted to evaluate the empirical performance of the proposed models. Finally, we illustrate the performance of the model with different applications.

E1452: **Robust differential variability testing for single-cell expression data: Bayes when single cells become big data**
Presenter: **Catalina Vallejos**, MRC Human Genetics Unit, United Kingdom

Cell-to-cell transcriptional variability in seemingly homogeneous cell populations plays a crucial role in tissue function and development. Single-cell RNA sequencing (scRNAseq) can characterise this variability in a transcriptome-wide manner. However, scRNAseq is prone to high levels of technical noise, creating new challenges for identifying genes that show genuine heterogeneous expression. We introduce BASiCS - a high-dimensional Bayesian hierarchical framework which simultaneously performs data normalisation, technical noise quantification and downstream differential expression analyses, whilst propagating statistical uncertainty across these steps. Beyond traditional mean expression testing, BASiCS can robustly identify changes in variability between cell populations, providing novel insights in e.g. immune cell populations. However, BASiCS was implemented using a MCMC algorithm which is time-consuming, particularly when applied to the large datasets that are increasingly available in scRNAseq experiments. We illustrate how recent scalable variations of MCMC as well as approximate inference methods can provide improvements in computational efficiency, and the associated trade-off with estimation performance. Finally, we also discuss some of the general challenges involved in applying these methods to the high-dimensional (Bayesian) models that are often required to capture the multiple sources of variability that underlie high-throughput omics data.

EO294 Room MAL G16 BAYESIAN INFERENCE VIA DISCRETE NONPARAMETRIC PRIORS Chair: Federico Camerlenghi

E0538: **A Berry-Esseen theorem for Pitman's alpha-diversity**
Presenter: **Stefano Favaro**, University of Torino and Collegio Carlo Alberto, Italy

The purpose is to study the random number K_n of blocks in the exchangeable random partition induced by a random sample of size n from the two parameter Poisson-Dirichlet process prior. Our main result is a Berry-Esseen theorem for the large n asymptotic behaviour of K_n. The proof relies on three intermediate novel results which may be of independent interest: i) a (probabilistic) representation of the distribution of K_n in terms of a compound Poisson distribution; ii) a quantitative version of the classical asymptotic expansion, in the sense of Poincaré, of a recurrent Laplace-type integral; iii) a refined quantitative bound for classical Poisson approximation. An application of our Berry-Esseen theorem is presented in the context of Bayesian nonparametric inference for species sampling problems, quantifying explicitly the error of a posterior approximation that has been extensively applied to infer the number of unseen species in a population.

E0554: **Bayesian inference for finite-dimensional discrete priors**
Presenter: **Tommaso Rigon**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Discrete random probability measures are the main ingredient for addressing Bayesian clustering. The investigation in this area has been very lively, with strong emphasis on nonparametric procedures based either on the Dirichlet process or on more flexible generalizations, such as the Pitman-Yor (PY) process or the normalized random measures with independent increments (NRMI). The literature on finite-dimensional discrete priors, beyond the classic Dirichlet-multinomial model, is much more limited. We aim at filling this gap by introducing novel classes of priors closely related to the PY process and NRMIs, which are recovered as limiting case. Prior and posterior distributional properties are extensively studied. Specifically, we identify the induced random partitions and determine explicit expressions of the associated urn schemes and of the posterior distributions. A detailed comparison with the (infinite-dimensional) PY and NRMIs is provided. Finally, we employ our proposal for mixture modeling, and we assess its performance over existing methods in the analysis of a real dataset.

E0721: **Universal boosting variational inference**
Presenter: **Trevor Campbell**, University of British Columbia, Canada

Boosting variational inference (BVI) approximates a probability density by building up a mixture of simple component distributions one at a time, using techniques from sparse convex optimization to provide both computational scalability and error guarantees. But the guarantees have strong conditions that do not often hold in practice, resulting in degenerate component optimization problems; and the ad-hoc regularization used to prevent degeneracy in practice can cause BVI to fail in unintuitive ways. The purpose is to introduce universal boosting variational inference (UBVI), a BVI scheme that exploits the simple geometry of probability densities under the Hellinger metric to prevent the degeneracy of other gradient-based BVI methods and avoid difficult joint optimizations of both component and weight. We will develop a scalable implementation of UBVI and show that for any target density and any mixture component family, the output converges to the best possible approximation in the mixture family, even when the mixture family is misspecified. We will discuss statistical benefits of the Hellinger distance as a variational objective through bounds on posterior probability, moment, and importance sampling errors. Experimental results will be provided, demonstrating that UBVI provides reliable and accurate posterior approximations with little to no tuning effort.

E1068: **Repulsive mixture modeling through Matern processes**
Presenter: **Vinayak Rao**, Purdue University, United States

Discrete mixtures are a flexible building block in the probabilistic modeler's toolkit, widely used in applications such as density estimation and clustering. An important issue arising from standard applications of discrete mixtures is low separation in the components; in particular, different components can be introduced that are very similar and hence redundant. Such a redundancy leads to extraneous clusters that are very similar, degrading performance, harming interpretability, and leading to computational problems and an unnecessarily complex models. Redundancy can arise in the absence of a penalty on components placed close together even when a Bayesian approach is used to learn the number of components.

To solve this problem, we propose a novel prior that generates components from a repulsive point process, viz. the Matern point process. Our model allows the number of mixture components to be estimated from data, automatically penalizing redundant components. We characterize this repulsive prior theoretically and propose an efficient Markov chain Monte Carlo sampling algorithm for posterior computation. The methods are illustrated using synthetic and real datasets.

EO430 Room CLO 101 REDUCTION TECHNIQUES FOR LARGE OR HIGH-DIMENSIONAL DATA Chair: Katja Ickstadt

E1197: **Sketches and coresets for large-scale statistical data analysis**
Presenter: **Alexander Munteanu**, TU Dortmund, Germany

The concepts of sketching and coresets are introduced. We will survey algorithmic techniques to construct those and show recent results in the context of linear and generalized regression problems.

E1221: **The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions**
Presenter: **Tamara Broderick**, MIT, United States

Discovering interaction effects on a response of interest is a fundamental problem faced in biology, medicine, economics, and many other scientific disciplines. In theory, Bayesian methods for discovering pairwise interactions enjoy many benefits such as coherent uncertainty quantification, the ability to incorporate background knowledge, and desirable shrinkage properties. In practice, however, Bayesian methods are often computationally intractable for even moderate-dimensional problems. Our key insight is that many hierarchical models of practical interest admit a particular Gaussian process (GP) representation; the GP allows us to capture the posterior with a vector of O(p) kernel hyper-parameters rather than $O(p^2)$ interactions and main effects. With the implicit representation, we can run Markov chain Monte Carlo (MCMC) over model hyper-parameters in time and memory linear in p per iteration. We focus on sparsity-inducing models and show on datasets with a variety of covariate behaviors that our method: (1) reduces runtime by orders of magnitude over naive applications of MCMC, (2) provides lower Type I and Type II error relative to state-of-the-art LASSO-based approaches, and (3) offers improved computational scaling in high dimensions relative to existing Bayesian and LASSO-based approaches.

E0464: **Unbiased estimators for linear regression and experimental design**
Presenter: **Michal Derezinski**, UC Berkeley, United States

Finding unbiased estimators for linear regression - where we wish to fit a linear function to a set of noisy measurements - is one of the oldest tasks in statistics. The classical Gauss-Markov theorem shows that the least squares estimator is the optimal solution for this problem under a set of strong assumptions regarding the data and measurement noise. Is it possible to construct an unbiased estimator for general random design linear regression without any assumptions on the measurement noise? We will show that it is possible by applying the least squares estimator to the dataset augmented by a small sample of additional measurements, generated from a certain determinantal point process called volume sampling. The obtained estimator is the first useful unbiased estimator for random design regression, and it can be efficiently constructed in many practical settings. As an example, we will show how this technique can be utilized in the context of A-optimal experimental design, where, given a large set of possible expensive measurements, we wish to select a small number of them to be performed, so as to construct an unbiased estimator with small mean squared error. Finally, we will discuss how these results extend to regularized least squares and Bayesian experimental design.

E0360: **Randomized methods for dimension reduction**
Presenter: **Benjamin Erichson**, ICSI and UC Berkeley, United States

In the era of Big Data, vast amounts of data are being collected and curated across the social, physical, engineering, biological, and ecological sciences. Techniques for dimensionality reduction, such as principal component analysis (PCA), are essential to the analysis of high-dimensional data. These methods take advantage of redundancies in the data in order to find low-rank, parsimonious models to reveal the underlying structure of the data. Classically, highly accurate deterministic matrix algorithms are used for this task. However, the emergence of large-scale datasets has severely challenged our computational ability to analyze data. Over the last decade, the concept of randomness has been demonstrated as an effective strategy to quickly produce approximate answers to familiar problems such as dimension reduction. Thus, the paradigm of randomized methods provides a scalable architecture for modern data science applications. These methods scale with the intrinsic rank of the data rather than the ambient dimensions of the measurement space. A brief overview of randomized methods for dimension reduction will be given.

EO729 Room Court SET-VALUED CLASSIFICATION Chair: Mohamed Hebiri

E2033: **Conformal prediction, a method that produces valid prediction sets under the assumption of exchangeability**
Presenter: **Lars Carlsson**, Centre for Machine Learning, Royal Holloway, University of London, United Kingdom

Conformal prediction can be used as a classification method in both supervised and unsupervised settings. It guarantees validity in the predictions under the exchangeability assumption. We will see how conformal prediction works for classification problems. Any machine-learning method can, with conformal prediction, produce predictions of label sets given a preset confidence level. The confidence level in predictions directly corresponds to the fraction of erroneous predictions made by the conformal predictor. This validity property will be demonstrated in an example. Furthermore, we will look at some different domains where conformal prediction has been successfully applied.

E1982: **Study of error rate reduction in mono-label classification using adaptive set-valued prediction**
Presenter: **Titouan Lorieul**, Zenith, LIRMM, University of Montpellier, Inria, France
Co-authors: Alexis Joly, Dennis Shasha

In presence of ambiguity in a multi-class classification task, usual single label predictions may be too limited. Set-valued predictions, on the other hand, provide answers similar to a human expert by allowing a predictor to output a set of candidate classes. Several formulations of this problem proposed in the literature result in an optimal strategy consisting in thresholding the regression function. We study the class of problems which would benefit most from such formulations of set-valued classification. In particular, the probability of error of the Bayes optimal classifier is compared to, (i) the best top-k classifier always predicting the k most probable classes, and, (ii) the previously mentioned optimal set-valued classifier resulting in an adaptive set size strategy. Conditions on the regression function quantifying the error rate reduction in the different cases are given. Experiments are carried out in order to test these assumptions on real-world datasets showing the usefulness of set-valued prediction in practice.

E0588: **Efficient algorithms for set-valued prediction in multi-class classification**
Presenter: **Thomas Mortier**, Ghent University, Belgium
Co-authors: Marek Wydmuch, Krzysztof Dembczynski, Eyke Hullermeier, Willem Waegeman

In cases of uncertainty, a multi-class classifier preferably returns a set of candidate classes instead of predicting a single class label with little guarantee. More precisely, the classifier should strive for an optimal balance between the correctness (the true class is among the candidates) and the precision (the candidates are not too many) of its prediction. We formalize this problem within a general decision-theoretic framework that unifies most of the existing work in this area. In this framework, uncertainty is quantified in terms of conditional class probabilities, and the quality of a predicted set is measured in terms of a utility function. We then address the problem of finding the Bayes-optimal prediction, i.e., the subset of class labels with highest expected utility. For this problem, which is computationally challenging as there are exponentially (in the number of classes) many predictions to choose from, we propose efficient algorithms that can be applied to a broad family of utility scores. Two of these

algorithms make use of structural information in the form of a class hierarchy, which is often available in prediction problems with many classes. Our theoretical results are complemented by experimental studies, in which we analyze the proposed algorithms in terms of predictive accuracy and runtime efficiency.

E1761: **Minimax semi-supervised confidence sets for multi-class classification**
Presenter: **Evgenii Chzhen**, Universite Paris-Est, France

Multiclass classification problems such as image annotation can involve a large number of classes. In this context, confusion between classes can occur, and a single label classification may fail. We will present a general device to build a confidence set classifier, instead of a single label classifier. In our framework the goal is to build the best confidence set classifier having a given expected size and the attractive feature of our approach is its semi-supervised nature - the construction of the confidence set classifier takes advantage of unlabeled data. The study of the minimax rates of convergence under the combination of the margin and non parametric assumptions reveals that there is NO supervised method that outperforms the proposed semi-supervised estimator. To further highlight the fundamental difference of supervised and semi-supervised methods, we establish that the best achievable rate for ANY supervised method is parametric, even if the margin assumption is extremely favourable. On the contrary, by using a sufficiently large unlabelled sample we are able to significantly improve this rate.

E1869: **A framework for online meta-learning**
Presenter: **Massimiliano Pontil**, Istituto Italiano di Tecnologia and University College London, Italy

The focus is on the problem in which a series of learning tasks are observed sequentially and the goal is to incrementally adapt a learning algorithm in order to improve its performance on future tasks. We consider both stochastic and adversarial settings. The algorithm may be parametrized by either a representation matrix applied to the raw inputs or by a bias vector. We develop a computational efficient meta-algorithm to incrementally adapt the learning algorithm after a task dataset is observed. The meta-algorithm performs online convex optimization on a proxy objective of the risk of the learning algorithm. We derive bounds on the performance of the meta-algorithm, measured by either the average risk of the learning algorithm on random tasks from the environment or by an average regret bound. Our analysis leverages ideas from multitask learning and learning-to-learn with tools from online learning and stochastic optimization. Lastly, we discuss extensions of the framework to nonlinear models such a deep neural nets and draw links between meta-learning, bilevel optimization and gradient-based hyperparameter optimization. A framework for online meta-learning

EO110 Room Jessel MISCELLANEOUS RESULTS ON DETECTION OF CHANGES Chair: Marie Huskova

E0380: **Multiple change-point detection in regression models via U-statistic type processes**
Presenter: **William Pouliot**, University of Birmingham, United Kingdom
Co-authors: Shixuan Wang

Many statistical and econometric procedures have been developed that are suited to testing for multiple changes in parameters of regression models which may occur at unknown times. Techniques have been developed or extended, but said extensions lack power for detecting changes in the intercept of linear regression models. A stochastic process that easily accommodates testing for many change-points that occur at unknown times has also been developed. It is shown via simulation that this U-statistic based processes lack power in finite samples for detecting change-points, even though the consistency of said tests has been established. A slight modification of his process is suggested which corrects for this problem. This slightly altered process is then used to fashion statistics which can be used to construct tests to detect multiple changes in intercept or variance of linear regression models, and will do so with much higher power than the original process. It is also shown that this slightly altered process, when weighted by appropriately chosen functions, is sensitive to detection of multiple changes in intercept that occur both early and later on in the sample, while maintaining sensitivity to changes that occur in the middle of the sample.

E0696: **High-dimensional changepoint detection via a geometrically inspired mapping**
Presenter: **Thomas Grundy**, STOR-i Centre for Doctoral Training, Lancaster University, United Kingdom
Co-authors: Rebecca Killick, Gueorgui Mihaylov, Jeremy Bradley

High-dimensional changepoint analysis is a growing area of research and has applications in a wide range of fields. The aim is to accurately and efficiently detect changepoints in time series data when the number of time points and dimensions grows large. Existing methods typically aggregate or project the data to a smaller number of dimensions; usually one. We present a high-dimensional changepoint detection method that takes inspiration from geometry to map the high-dimensional time series to two dimensions. Applying univariate changepoint detection methods to both mapped series allows the detection of changepoints that correspond to changes in the mean and variance of the original time series. We demonstrate that this approach outperforms the current state-of-the-art multivariate changepoint methods both in the accuracy of detected changepoints and computational efficiency.

E0884: **Testing changed segment by maximal ratio statistics**
Presenter: **Alfredas Rackauskas**, Vilnius University, Lithuania

A new test statistic is proposed for testing changed segment in a sample of regularly varying random variables. The basic idea is to take maximal ratios of weighted moving sums, avoiding possible normalization problems and, at the same time, providing tools for testing small changed segments. We show how the asymptotic distribution of the maximal ratio statistics depends on weights of moving sums and provide some empirical illustrations.

E0911: **Testing and estimation of change-points in covariance matrices based on weighted CUSUMs in high-dimensions**
Presenter: **Ansgar Steland**, University Aachen, Germany

The analysis of high-dimensional covariance matrices of time series is a challenging statistical problem. We approach the problem to test for the presence of a change-point in a sequence of covariance matrices by studying procedures based on weighted CUSUM statistics associated to bilinear forms of the sample covariance matrix. Asymptotic results in terms of strong and weak approximations as well as functional central limit theorems are presented under a change-point time series model. We consider a linear time series framework which allows for approximate VARMA models and a class of spiked covariance models. Further, the results cover approximations for a multivariate CUSUM transform based on L pairs of projection vectors. Consistent estimators for the asymptotic variances and covariances of the weighted CUSUM statistics are considered. Studying sequential versions of these estimators allows us to consider a stopped-sample estimator which uses the data up to the estimated change point. The finite sample properties are investigated by simulations. Lastly, the methods are illustrated by analyzing environmental data.

EO709 Room MAL 152 RESAMPLING AND SIMULATIONS FOR INFERENCE IN COMPLEX SETTINGS Chair: Maria-Pia Victoria-Feser

E0321: **Simulated switched Z-estimation for accurate finite sample inference**
Presenter: **Samuel Orso**, University of Geneva, Switzerland
Co-authors: Maria-Pia Victoria-Feser, Stephane Guerrier, Mucyo Karemera

Constructing tests or confidence regions that control over the error rates in the long-run is probably one of the most important problem in statistics. Yet, the theoretical justification for most methods in statistics is asymptotic. The bootstrap for example, despite its simplicity and its widespread usage, is an asymptotic method. There is, in general, no claim about the exactness of inferential procedures in finite sample. We propose an

alternative to the parametric bootstrap. We set up general conditions to demonstrate theoretically that accurate inference can be claimed in finite sample.

E0402: **Right-censoring bias correction for growth curve linear mixed models**
Presenter: **Dominique-Laurent Couturier**, University of Cambridge, United Kingdom
Co-authors: Stephane Guerrier, Maria-Pia Victoria-Feser

Tumour growth inhibition studies typically involve analysing tumour sizes measured regularly over a period of time. The aim is usually to detect differences in growth rate between experimental conditions. Many methods have been considered. Some summarise each growth curve into a single measure and compare the location parameter of these statistics between different experimental conditions by means of Welsh tests. Others consider mixed/longitudinal models, taking into account the time and within tumour dependence of the observations to provide a parametric fit on all collected data. As animals are culled when their tumour size exceeds a legal upper limit or when the discomfort level is considered too high, such data are often right censored, leading to biased growth estimates. The objective is to develop a method allowing one to correct the bias of growth curve linear mixed models in the presence of right censoring due to a fixed upper tumour size limit. Simulations show that the iterative bootstrap bias corrected estimator we developed for random intercept and slope mixed models allows us to obtain unbiased growth rate estimates as well as confidence intervals showing coverages close to the nominal value.

E0868: **Finite sample unbiased estimation in high dimensional settings**
Presenter: **Mucyo Karemera**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Samuel Orso, Maria-Pia Victoria-Feser

Considering the increasing size of available data, the need for statistical methods that control the finite sample bias is growing. This is mainly due to the frequent settings where the number of variables is large and allowed to increase with the sample size bringing standard inferential procedures to incur significant loss in terms of performance. Moreover, the complexity of statistical models is also increasing thereby entailing important computational challenges in constructing new estimators or in implementing classical ones. A trade-off between numerical complexity (e.g. approximations of the likelihood function) and statistical properties is often accepted. However, numerically efficient estimators that are altogether unbiased in finite sample, consistent and asymptotically normal in high-dimensional problems would be advantageous, especially for real data applications. We set a general framework from which such estimators can be easily derived for wide classes of models. The approach allows various extensions compared to previous results as it is adapted to possibly inconsistent estimators and is applicable to discrete models and/or models with a large number of parameters (compared to the sample size).

E1227: **Exact finite sample inference for studies with a small number of clusters**
Presenter: **Stephane Heritier**, Monash University, Australia
Co-authors: Maria-Pia Victoria-Feser, Stephane Guerrier

Cluster randomised trials (CRTs), particularly longitudinal CRTs, often generate data with a small number of clusters typically analysed using generalised estimating equations (GEE) or generalised mixed models. A major drawback of standard techniques is that they are asymptotic in nature and rely on a large number of clusters to be valid. Ignoring this leads to: 1) a grossly-inflated type 1 error or, in general, confidence intervals (CIs) that are too short; 2) biased variance or intra-cluster correlation (ICC) estimates. We propose a new simulation-based approach allowing exact finite-sample inference for such problems. The idea is to compute first an initial simple(r) estimator, possibly biased, of the parameter of interest. In a second step, simulations are used to correct the bias via a novel algorithm called the Iterative Bootstrap (IB). The finite sample distribution can be generated under weak regularity conditions. We study the performance of this approach by simulations in various settings and show that it outperforms standard methods based on asymptotic sandwich variance formula with/without small sample correction. This indirect approach shows extremely promising results both theoretically and empirically.

EO324 Room MAL 153 NON-REGULAR STATISTICAL MODELING AND COMPUTATIONAL METHODS Chair: Tsung-I Lin

E0567: **An improved approach for estimating large losses using the g-and-h distribution**
Presenter: **Marco Bee**, University of Trento, Italy
Co-authors: Luca Trapin, Julien Hambuckers

The g-and-h distribution finds applications in modeling highly skewed and fat-tailed data, like extreme losses in the banking and insurance sector. Given the lack of a closed-form density, two estimation methods are introduced: a maximum likelihood technique based on a numerical approximation of the likelihood function, and an indirect inference approach with a bootstrap weighting scheme. A realistic simulation study suggests that indirect inference is computationally more efficient and provides better estimates in case of extreme features of the data, whereas maximum likelihood is preferable in terms of root-mean-squared-error when the data are less skewed and heavy-tailed. Empirical illustrations on insurance and operational losses illustrate these findings.

E0853: **Analysis of interval data using patterned covariance structures**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States

Principal component analysis of interval data is proposed by using block compound symmetry (BCS) and doubly block compound symmetry (DBCS) covariance structures. This is deemed by considering each interval as two repeated measurements at the lower and upper bounds of the interval (two-level multivariate data), and then by assuming BCS covariance structure for the data. And, this is accomplished in two stages: first getting eigenblocks and eigenmatrices of the variance-covariance matrix, and then analyzing these eigenblocks and the corresponding principal vectors together to get the adjusted eigenvalues and the corresponding eigenvectors of the interval data. We then work independently with these principal vectors and their corresponding variance-covariance matrices, i.e., the corresponding eigenblocks to get the eigenvalues and eigenvectors of the interval data. If there is some additional information (like brands etc.) in the interval data, the interval data can be considered as three-level multivariate data and can be analyzed by assuming DBCS covariance structure. Results illustrating the appropriateness of the new methods over the existing methods are presented. It is shown that our proposed method of principal component analysis for three-level interval data generalizes the commonly used PCA for multivariate data. The proposed methods is illustrated with a real dataset.

E1465: **Likelihood-based inference for mixed-effects models with censored response using the skew-normal distribution**
Presenter: **Victor Hugo Lachos Davila**, University of Connecticut, United States
Co-authors: Larissa Avila Matos, Thalita do Bem Mattos

Mixed-effects models are commonly used to fit longitudinal or repeated measures data. A complication arises when the response is censored, for example, due to limits of quantification of the assay used. Although normal distributions are commonly assumed for random effects, such an assumption may be unrealistic obscuring important features of among-individual variation. We relax this assumption by consider a likelihood-based inference for linear and nonlinear mixed effects models with censored response (NLMEC/LMEC) based on the multivariate skew-normal distribution. An ECM algorithm is developed for computing the maximum likelihood estimates for NLMEC/LMEC with the standard errors of the fixed effects and the exact likelihood value as a by-product. The algorithm uses closed-form expressions at the E-step, that rely on formulas for the mean and variance of a truncated multivariate skew-normal distribution. The proposed algorithm is implemented in the R package skewlmec. It is applied to analyze longitudinal HIV viral load data in two recent AIDS studies. In addition, a simulation study is conducted to examine the performance of the proposed methods.

E0676: **Family of matrix-variate distributions: A flexible approach based on the mean-mixture of normal model**
Presenter: **Mehrdad Naderi**, University of Pretoria, South Africa
Co-authors: Andriette Bekker
A new family of matrix-variate distributions is introduced which is based on the mean-mixture of normal (MMN) model. The properties of the new matrix-variate family, namely, stochastic representation, moments and characteristic function, linear and quadratic forms as well as marginal, conditional distributions, are investigated. Three special cases including the restricted skew-normal, exponentiated MMN and the mixed-Weibull MMN matrix variate distributions are presented and studied. Maximum likelihood estimate of the parameters are obtained by implementing an EM-type algorithm. The usefulness and practical utility of the proposed methodology are illustrated through two conducted simulation studies and through the landsat satellite dataset.

EO284 Room MAL 254 HIGHLY STRUCTURED STOCHASTIC SYSTEMS Chair: Marie-Colette Van Lieshout

E0245: **Markov object processes: From area-interaction to linear networks**
Presenter: **Marie-Colette Van Lieshout**, CWI/UT, Netherlands
An overview of Markov object processes and their applications in image analysis will be given starting from the pioneering work in the 1980s on segmentation by means of Markov random fields. Influence zone based spatio-temporal point processes, deformable template models and sequential object processes will be presented. Such models are useful for higher level image analysis tasks including tracking and recognition. Turning to the intermediate level, we will discuss polygonal Markov field models and show that, on the one hand, discrete versions of such mosaics are dual to Markov random fields, and, in the other direction, that certain polygonal Markov field models can be seen as a non-overlapping marked point process. Finally, we will briefly indicate how to construct point processes on linear networks.

E0513: **Semi-parametric multinomial logistic regression for multivariate point processes**
Presenter: **Rasmus Waagepetersen**, Aalborg University, Denmark
Multivariate point pattern data are becoming increasingly common. In ecology for example, biologists collect large data sets of locations of hundred thousands trees belonging to hundreds of species. Similarly, in many major cities, police authorities record locations, times and types of street crimes. We discuss a semi-parametric approach to analysing street crime data where the intensity functions of different types of crime scenes are specified by regression models up to a common unknown spatially varying factor. This factor may e.g. represent variations in crime intensity due to complex urban structures and population density. No restrictive assumptions of independence within or between crime types are imposed. We discuss how inference on the intensity functions can be conducted using a multinomial conditional composite likelihood. In this connection we address how to estimate standard errors of the regression parameter estimates where these standard errors depend on the multivariate dependence structure between the different types of points. We apply the methodology to a data set of street crimes from Washington DC and show how interesting spatial patterns emerge as a result of our analysis.

E1042: **Modelling central place foraging of wolves**
Presenter: **Juha Heikkinen**, Natural Resources Institute Finland (Luke), Finland
Co-authors: Anna-Kaisa Ylitalo, Ilpo Kojola
Central place foraging refers to a common movement behaviour of many wild animals, where the most notable regular feature are foraging trips starting from and ending to a fixed residence. Such movement is particularly typical to adult wolves during the early summer. The pups are then still too small to join in foraging. They stay near the den, and the adults return regularly to feed them. We present a hidden Markov model (HMM), which captures the main movement modes related to the different parts of the typical foraging trips of wolves. The model was developed on the basis of twelve intensively monitored two months movement tracks with approximately 2,500 GPS-relocations per track. In model validation, the focus is on the ultimate aims. First, simulations from the null model of independent trips are required to evaluate different hypotheses concerning possible dependence between subsequent trips. Secondly, the HMM should be a useful component in the interpolation model, which would allow us to utilize the 200 less intensively monitored tracks more efficiently.

E1374: **Conditional independence and the Gaussian distribution**
Presenter: **Haavard Rue**, KAUST, Saudi Arabia
An important focus in the HSSS was the ability to build statistical models from smaller building blocks, either through the directed acyclic graph (through WinBUGS that was introduced about that time), or using Markov random fields (MRFs) and conditional independence. Due to relative newly (re-)invented Gibbs sampling algorithm and relatives, then Bayesian inference could be conducted using MCMC. We will discuss the Gaussian case and argue for why conditional independence is such an important concept there. Also, why it is important from a computational point of view, as it allows us to factorise very large sparse matrices, which is useful also for approximate Bayesian inference.

EO108 Room Senate WEATHER AND CLIMATE EXTREMES Chair: Marco Oesting

E1165: **Sub-daily rainfall extremes in Australia: An analysis of space-time variability and drivers across the continent**
Presenter: **Kirstin Strokorb**, Cardiff University, United Kingdom
Co-authors: Marie Ekstrom, Owen Jones, Aidan Gibbons, Hayley Fowler
Heavy precipitation events may have severe and harmful impacts on the environment for (human) life. On a global scale, a warming climate suggests an increased moisture holding capacity of the atmosphere and thereby an intensification of heavy rainfall events. While current understanding of temporal and spatial variability in rainfall extremes is mostly based on observed daily accumulations, reported changes in sub-daily extreme events (flash-floods) are much less well understood. We use 1-hour accumulation observations collected via the INTENSE project and a Bayesian hierarchical model to investigate regional characteristics and the role of geophysical drivers of rainfall extremes in Australia with a particular emphasis on 3-hourly totals.

E0943: **A hierarchical max-infinitely divisible process for extreme areal precipitation over watersheds**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Benjamin Shaby, Gregory Bopp
Understanding the spatial extent of extreme precipitation is necessary for determining flood risk and adequately designing infrastructure (e.g., stormwater pipes) to withstand such hazards. While environmental phenomena typically exhibit weakening spatial dependence at increasingly extreme levels, limiting max-stable process models for block maxima have a rigid dependence structure that does not capture this type of behavior. In order to model block maxima at sub-asymptotic regimes, we suggest using models from a broader family of max-infinitely divisible (max-id) processes, which retain appealing properties reflecting the specific dependence structure of maxima, while allowing for weakening spatial dependence at increasingly high levels. We will first present general construction principles for max-id processes and discuss how flexible asymptotically independent max-id models may be designed. We will then describe a Bayesian max-id process, whose likelihood function admits a hierarchical representation in terms of random effects, and which scales well to large datasets. The proposed model is constructed using flexible random basis functions that are estimated from the data, allowing for straightforward inspection of the predominant spatial patterns of extremes. We apply our model to extreme precipitation in eastern North America, and show that the proposed model adequately captures the extremal behavior of the data.

E1526: **Trends in the extremes of environments associated with severe US thunderstorms**
Presenter: **Jonathan Koh**, EPFL, Switzerland
Co-authors: Erwan Koch, Anthony Davison
Concurrently high values of convective available potential energy (CAPE) and storm relative helicity (SRH) are conducive to hazardous convective weather associated with severe thunderstorms. Hence, it is highly relevant to have probabilistic models for both variables' extremes that use relevant covariate information to account appropriately for their spatial and temporal dependence. We consider a large area of the contiguous United States over the period 1979–2015 and use statistical extreme value models and appropriate multiple testing procedures to identify trends in the extremes. In the first step, we show that there is a significant time trend in the extremes for CAPE and SRH maxima in the spring and summer months. These increases in CAPE are also relevant for rainfall extremes and are expected in a warmer climate but have not previously been reported. In the second step, we focus on the local spatial extremal dependence structure and thus model the pointwise maxima using max-stable random fields. We focus especially on the Brown–Resnick field with a power variogram and investigate the effect of the El Niño-Southern Oscillation (ENSO) on its parameters with the use of tensor product splines. Our results show that the range parameters for CAPE and SRH are lower in the spring and summer, so the corresponding extremes are more localized during these seasons than in the winter. Lastly, we find that these seasonal differences are more pronounced during La Niña events.

E1248: **Robust estimation of the extremal index in the context of climate time series**
Presenter: **Katharina Hees**, TU Dortmund University, Germany
Extreme events of climate time series often occur in clusters, for example, in storms, floods, earthquakes, etc. The most common approach to analyze such serially correlated data is to first identify the clusters and then to proceed with the peaks of the clusters with classical extreme value theory methods. The extremal index θ plays an important role in the declustering process. One interpretation of this quantity is, for example, that it is the reciprocal of the limiting mean number of exceedances in blocks with at least one exceedance. Another interpretation is that it is the proportion of inter-exceedance times that represent the times between different clusters. Hence, the knowledge of the extremal index allows us to decluster the data by sorting the inter-exceedance times by size and assuming the θ largest to be the intercluster times (between clusters) and the $(1-\theta)$ smallest to be the intracluster times (within clusters). In the context of climate time series, one is often confronted with outliers. Several methods for the estimation of the extremal index were proposed in the literature, but most of them are not robust with respect to such outliers. We will present a method that is robust, and compare it to existing and well-established extremal index estimators.

EO524 Room MAL 253 RECENT DEVELOPMENTS IN THE ANALYSIS OF NEUROIMAGING AND GENETIC DATA Chair: Farouk Nathoo

E1553: **A Bayesian spatial model for imaging genetics**
Presenter: **Liangliang Wang**, Simon Fraser University, Canada
Co-authors: Yin Song, Shufei Ge, Jiguo Cao, Farouk Nathoo
A Bayesian bivariate spatial group lasso model is developed for multivariate regression analysis applicable to studies examining the influence of genetic variation on brain structure. The model is motivated by an imaging genetics study of the Alzheimer's Disease Neuroimaging Initiative (ADNI), where the objective is to examine the association between images of volumetric and cortical thickness values summarizing the structure of the brain as measured by magnetic resonance imaging (MRI) and a set of 486 SNPs from 33 Alzheimer's Disease (AD) candidate genes obtained from 632 subjects. A bivariate spatial process model is developed to accommodate the correlation structures typically seen in structural brain imaging data. First, we allow for spatial correlation in the imaging phenotypes obtained from neighbouring regions on the same hemisphere of the brain. Second, we allow for correlation in the same phenotypes obtained from different hemispheres (left/right) of the brain. We develop a mean-field variational Bayes algorithm and a Gibbs sampling algorithm to fit the model. We also incorporate Bayesian false discovery rate (FDR) procedures to select SNPs. We implement the methodology in a new release of the R package bgsmtr. We show that the new spatial model demonstrates superior performance over a standard model in the motivating application.

E1566: **Empowering association tests using unpaired data**
Presenter: **Kayhan Batmanghelich**, University of Pittsburgh, United States
There is a growing interest in the biomedical research community to incorporate retrospective data, available in healthcare systems, to shed light on associations between different biomarkers. Understanding the association between various types of biomedical data, such as genetic, blood biomarkers, imaging, etc provides a holistic understanding of human diseases. To test the association hypothesis between two types of data in Electronic Health Records, one requires a substantial sample size with both data modalities to achieve a reasonable power. Current methods only allow using data from individuals who have both data modalities. Hence, researchers cannot take advantage of much larger samples in EHR that have at least one of the data types, which limits the power of the association test. We present a new method called the Semi-paired Association Test (SAT) that makes use of both paired and unpaired data. In contrast to classical approaches, incorporating unpaired data allows the SAT to produce better control of false discovery and to improve the power of the association test. We study the properties of the new test theoretically and empirically, through a series of simulations and by applying our method on real studies in the context of chronic disease. We are able to identify an association between the high-dimensional characterization of CT chest images and several blood biomarkers as well as the expression of dozen of genes involved in the immune system.

E1631: **Hierarchical Bayesian mixture modeling of resting-state functional brain connectivity: An alternative to thresholding**
Presenter: **Tetiana Gorbach**, Umea University, Sweden
Co-authors: Anders Lundquist, Xavier de Luna, Lars Nyberg, Alireza Salami
A Bayesian hierarchical mixture model is proposed in order to analyze functional brain connectivity where mixture components represent connected and non-connected brain regions. Such an approach provides a data-informed separation of reliable connections from noise in contrast to arbitrary thresholding of a connectivity matrix. The hierarchical structure of the model allows simultaneous inferences for the entire population and each subject separately. We show that a new connectivity measure, the posterior probability of a given pair of brain regions of a specific subject to be connected given the observed correlation of regions activity, might be superior to correlation. The posterior probability reflects connectivity of a pair of regions relative to the overall connectivity pattern of an individual, which is overlooked in traditional correlation analyses. We also demonstrate that using the posterior probability might diminish the effect of noise on inferences, which is present when a correlation is used as a connectivity measure. Additionally, simulation analyses reveal that the sparsification of the connectivity matrix using the posterior probabilities might outperform the absolute thresholding based on correlations. The applicability of the introduced method is exemplified by a study of functional resting-state brain connectivity in older adults.

E1908: **Clustering of generalized gamma distributions by using information geometry: An application to medical imaging**
Presenter: **Florence Nicol**, ENAC, France
Co-authors: Sana Rebbah, Stephane Puechmorel
Probability density functions can be treated as functional data and may be represented as points of a statistical manifold using Information Geometry. Within this frame, densities are endowed with a Riemannian manifold structure, the metric being generally given by the Fisher information. The purpose is to present some new results about the generalized gamma manifold and how information geometry improved the performance of the classification of Alzheimer's disease population. In the medical field, a growing number of quantitative image analysis techniques have been developed, including analysis of histograms, which is widely used to quantify the diffuse pathological changes of some neurological diseases. For

using the entire information included in the data, the underlying probability density functions themselves should be rather used as a biomarker of the whole brain. Some information geometric properties of the generalized gamma family are investigated, especially when restricted to the gamma submanifold, that is particularly relevant in the Alzeihmers disease context. The Fisher information and results in the case of the generalized gamma manifold will be first detailed. Next, a clustering technique has been successfully extended by using a geodesic distance of which an approximation is computed numerically with a two steps algorithm.

EO366 Room SH349 STATISTICAL METHODS FOR SPORTS Chair: Ivor Cribben

E0191: **From pixels to points: Using tracking data to measure performance in professional sports**
Presenter: **Luke Bornn**, Simon Fraser University, Canada

The aim is to explore how players perform, both individually and as a team, on a basketball court. By blending advanced spatio-temporal models with geography-inspired mapping tools, we are able to understand player skill far better than either individual tool allows. Using optical tracking data consisting of hundreds of millions of observations, we will demonstrate these ideas by characterizing defensive skill and decision making in NBA players.

E0274: **A deeper understanding of quarterback pressure in football**
Presenter: **Karl Pazdernik**, Deep Football, United States
Co-authors: Jacques Kvam

Quarterback pressure is a key component of any good defense in football. Yet, the very idea of pressure is not well quantified. Current metrics, such as hurries, knockdowns, and sacks, are useful, but they are binary outcomes and are incomplete in how they assign credit and fault. For example, a defender coming close to the quarterback is usually blamed on the offensive lineman, but could be because the quarterback held the ball too long or ineffectively used the space provided. Using a Voronoi tessellation applied to RFID tracking data, we measure the space of the pocket and calculate the pole of inaccessibility, i.e. the best location for the quarterback to minimize pressure. We then use these measures over time to attribute credit and fault to all contributing defensive and offensive players, respectively. The result is a more thorough assessment of rushing defenders, pass-blocking offensive players, and the quarterbacks ability to reduce pressure himself.

E1504: **The hot hand theory in hockey: A multilevel logistic regression analysis**
Presenter: **Armann Ingolfsson**, University of Alberta, Canada
Co-authors: Ivor Cribben, Likang Ding

The Hot Hand theory states that an athlete will perform better in the present if he/she has performed well in the recent past. This theory has been investigated for basketball, baseball, and other sports. We test this theory for National Hockey League (NHL) playoff goaltenders by estimating how their performance on recent shots influences the probability of saving the next shot on goal. We use multilevel logistic regression models, in which we allow either some coefficients to vary among the season-goaltender combinations. In our regression model, the recent performance of a goaltender is measured by the number of saves within the most recent shots he faced during the same game. We also include other control variables such as shot type, shot origin, and game score in our model. Our data consists of 36,235 shot-on-goal observations for 90 goaltenders who played in the NHL playoffs between 2008 and 2016. We compare the results of multilevel models to simple logistic models as well as to multilevel models with different settings. Our preliminary findings are that a good recent save performance has a negative effect on the save probability for the next shot, which is consistent with the opposite of the Hot Hand theory.

E1816: **Markov decision processes in sports analytics**
Presenter: **Oliver Schulte**, Simon Fraser University, Canada

Markov decision processes are a fundamental framework for optimizing sequential decisions. We describe how they can be applied in sports analytics, to provide a powerful statistical analysis of dynamic sports data. The main idea is to build a Markov decision process model of a sport and estimate a value function for it using reinforcement learning. Large-scale models have been built for several sports, including hockey, soccer, golf, basketball, and American football. The focus is on player evaluation, a fundamental problem of sports analytics. We will describe a neural net model of a value function trained on over 3M play-by-play events in the National Hockey League, the leading ice hockey league. We give several natural definitions of player performance derived from the value function, and prove their equivalence. Empirical evaluation shows that the resulting player ranking is consistent throughout a play season, and correlates highly with standard success measures and future salary.

EG836 Room MAL 251 CONTRIBUTIONS IN RESTRICTED PARAMETERS INFERENCE AND SHRINKAGE Chair: Eric Marchand

E1711: **On some beta ridge regression estimators: Methods, simulation and application**
Presenter: **Muhammad Qasim**, Jonkoping University, Sweden
Co-authors: Kristofer Mansson, BM Golam Kibria

The classic statistical method for modelling the rates and proportions is the beta regression model (BRM). The BRM is applicable when the dependent variable is continuous, beta distributed and limited in the interval (0, 1). The standard maximum likelihood estimator (MLE) is used to estimate the coefficients of the BRM. However, this MLE is very sensitive when the regressors are linearly correlated to each other. Therefore, a new beta ridge regression (BRR) estimator is introduced as a remedy to the problem of instability of the MLE. We study the mean square error properties (MSE) of this estimator analytically, and then, based on the derived MSE, we suggest some new estimators of the shrinkage parameter. We also suggest a median square error (SE) performance criteria which can be used to achieve strong evidence in favor of proposed method for the Monte Carlo simulation study. The performance of BRR and MLE is appraised by means of Monte Carlo simulation where mean and median SE are used as performance criteria. We found that the proposed estimators performed better than some existing estimators. Finally, an empirical application is used to show the advantages of the proposed estimator.

E1786: **Improved estimators for zero-inflated count data models in the presence of multicollinearity**
Presenter: **Talha Omer**, Jonkoping University, Sweden, Sweden
Co-authors: Par Sjolander, Kristofer Mansson, BM Golam Kibria

Zero inflated count-data models are used when the data is in the form of non-negative integers. A surplus of zeros induces overdispersion in the dependent variable of the count regression model. Under these circumstances, zero-inflated models can be used effectively. There is a clear empirical relevance for these types of models, for instance when modelling the demand for health services when most patients have zero visits, or when counting the number of insurance claims within population, etc. However, multicollinearity is a frequently observed, but usually disregarded, empirical problem for these types of data sets. Multicollinearity increases the variance of the estimated coefficients and make the estimates very sensitive. Therefore, we address this relevant problem by considering some improved estimators such as Ridge and Liu estimators for non-negative count models. The performance of these estimators has been evaluated by Monte Carlo simulations, and based on the MSE and the MAE performance criteria, the simulations illustrate that our improved estimators better than the usual maximum likelihood estimator and some other Liu estimators in the presence of multicollinearity. At the end, an empirical application is conducted for the improved Liu and ridge estimators and its results support the simulation study.

E1741: **Adapting the horseshoe prior for functional effects in distributional regression models**
Presenter: **Paul Wiemann**, University of Goettingen, Germany

Co-authors: Thomas Kneib

A new prior specification based on the horseshoe prior is proposed that allows us to carry the concept of Bayesian global-local shrinkage to functional effect types in the class of distributional regression models. Distributional regression models link structured additive predictors to every distributional parameter via a response function. These predictors can be composed of various effect types, e.g., non-linear effects, varying coefficients random effects, spatial effects and may include hierarchical regression structures. The presented approach adaptively shrinks the estimated effect towards a predefined functional subspace, i.e. a linear function, while keeping desirable properties of the horseshoe prior unchanged, namely, those concerning the handling of sparsity and adaptive shrinkage. A Markov chain Monte Carlo sampling scheme is provided. Using simulated data and real data, empirical experiments show that our approach is applicable in situations with a large number of covariates and non-normal response distributions.

E1784: **A comparison of preliminary test, stein-type and penalty estimators in gamma regression models**

Presenter: **Akram Mahmoudi**, Jonkoping International Business School, Sweden

Co-authors: Reza Arabi Belaghi, Saumen Mandal

Some estimators are proposed based on the preliminary test and Stein-type strategies to estimate unknown parameters in a Gamma regression model. These proposed estimators are considered when it is suspected that the parameters may be restricted to a subspace of the parameter space. Also, two penalty estimators such as LASSO and ridge regression are presented. Comprehensive Monte-Carlo simulation experiments are conducted. Then, comparisons are made based on simulated relative efficiency to clarify the performance of the proposed estimators. Practitioners are recommended to use the positive-part Stein-type estimator since its performance is robust irrespective of the reliability of the subspace information. A real data on prostate cancer is considered to illustrate the performance of the proposed estimators.

CO230 Room MAL B04 SPATIAL INEQUALITIES: MEASUREMENTS AND METHODS Chair: Marzia Freo

C0449: **Dynamics of spatial autocorrelation: A new space-time state of mind**

Presenter: **Roberto Patuelli**, University of Bologna, Italy

Co-authors: Solmaria Halleck-Vega

The dynamics of spatial autocorrelation, which is present in most spatial data, are explored. Usually it is accounted for either in the error term and/or using other spatial econometric and statistic techniques. It can thus either be treated as a nuisance or substantive phenomenon. In contrast to studies focusing on including spatial effects in regression models, we explore actual changes in spatial autocorrelation over time. This offers a distinctive perspective. As a first step, it is useful to appraise the dynamics and relationships between time series data and the respective evolution of spatial autocorrelation. Then, to better understand whether trend-cycle and/or seasonal components have a role to play in explaining spatial autocorrelation, a time series decomposition can be applied. As a third step, it can be further explored if dynamics of spatial autocorrelation coincide with relevant factors such as favourable or unfavourable socio-economic trends and policy changes. We highlight different regional labor markets in Europe, which makes for an interesting exploration due to their diversity and policy relevance.

C1629: **Estimation of area-wise spatial income distributions from grouped data**

Presenter: **Shonosuke Sugasawa**, University of Tokyo, Japan

Co-authors: Genya Kobayashi, Yuki Kawakubo

Estimating income distributions plays an important role in the measurement of inequality and poverty over space. The existing literature on income distributions predominantly focuses on estimating an income distribution for a country or a region separately and the simultaneous estimation of multiple income distributions has not been discussed in spite of its practical importance. We develop an effective method for the simultaneous estimation and inference for area-wise spatial income distributions taking account of geographical information from grouped data. Based on the multinomial likelihood function for grouped data, we propose a spatial state-space model for area-wise parameters of parametric income distributions. We provide an efficient Bayesian approach for estimation and inference for area-wise latent parameters, which enables us to compute area-wise summary measures of income distributions such as mean incomes and Gini indices, not only for sampled areas but also for areas without any samples thanks to the latent spatial state-space structure. The proposed method is demonstrated using the Japanese municipality-wise grouped income data. The simulation studies show the superiority of the proposed method to a crude conventional approach which estimates the income distributions separately.

C0251: **Entropy as measure of spatial agglomeration: Interactions of business locations and housing transactions**

Presenter: **Katarzyna Kopczewska**, University of Warsaw, Poland

Entropy is usually used in measuring the concentration of business (often called specialisation) and applied to data aggregated by sectors and regions. This traditional approach does not respond spatial analyses challenges and should be extended for point data to measure the spatial agglomeration over space. This however requires transformation of data to obtain the weights of points. Voronoi tessellation tiles can approximate well the point pattern in the continuous space. The switch from point data to polygonal representation also changes the approach to spatial relation, mainly distance between objects. Instead of measuring the spatial separation between points, one can use the share of tiles surface in whole area to conclude about the closest neighbourhoods. This also opens the opportunities of using entropy, because of proportional character of percentage areal data. Thus, the tessellated point pattern can be examined for the existence of agglomeration with entropy measure. Comparative analyses on the location patterns and density of points are of particular interest in urban studies. However, aggregated data for urban areas erase many of spatial information, what lowers the analysis power in this very highly diversified environment. Urban studies benefit from low-granulation data, especially point data. Point data applied to entropy via tessellation can be used to understand how the spatial allocation interacts and to detect spatial agglomeration of point data.

C1771: **Estimating the heterogeneous impact of the EU cohesion policy across regions**

Presenter: **Marzia Freo**, University of Bologna, Italy

Co-authors: Elena Calegari, Aura Reggiani

The Cohesion Policy (CP) represents the main territorial policy of the European Union, aiming to reduce disparities among regions. To address this aim, the largest majority of the funds are devoted to the poorest regions, called Objective 1 regions. During the last years, in response to the debate on the effectiveness of the CP, a broad empirical literature on the impact evaluation of the EU funds has flourished. In the debate, some consensus on the effectiveness of the EU transfers emerges, but with additional distinctions; for instance, various researches focus on the heterogeneity of the effect of CP funds across European regions. A further contribution to this strand of literature is provided by estimating the different quantiles of the effect of CP on regions. More specifically, the impact of the CP on the regional per capita GDP growth is analysed. The applied methodology allows to estimate the counterfactual distribution of the regional GDP growth that Objective 1 regions would have attained in absence of the CP funds. By comparing this estimated distribution with the observed one, it is possible to obtain the whole distribution of the policy impact. The results, related to the programming period 2007-2013, confirm the positive impact of CP on regional economic growth with a larger impact on less performing regions.

CO168 Room MAL B20 FRACTIONAL MOTIONS AND ARTIFICIAL NEURAL NETWORKS FOR TIME SERIES Chair: Matthieu Garcin

C1730: **A distribution-based method to gauge market liquidity through scale invariance between investment horizons**

Presenter: **Sergio Bianchi**, University of Cassino and Southern Lazio, Italy

Co-authors: Augusto Pianese, Massimiliano Frezza

A new nonparametric and distribution-based method is developed to detect self-similarity among the rescaled distributions of the log-price variations over a number of time scales. The procedure allows us to test the statistical significance of the scaling exponent that possibly characterizes each pair of time scales, and to study the link between self-similarity and liquidity, the core assumption of the Fractal Market Hypothesis (FMH). The method can support financial operators in the selection of the investment horizons to be preferred as well as regulators in the adoption of guidelines to ensure the stability of markets. The analysis performed on the S&P500 reveals a very complex, time-changing scaling structure, which confirms the link between market liquidity and self-similarity.

C1743: **Multiscaling in finance**

Presenter: **Tiziana Di Matteo**, Kings College London, United Kingdom

The multiscaling behaviour of financial time series is one of the acknowledged stylized facts in the literature. The source of the measured multifractality in financial markets has been long debated and it has been attributed to mainly two sources: the power law tails and the non linear autocorrelation of the analysed time-series. We will discuss the origin of multiscaling in financial time-series and investigate how to best quantify it. In particular, we will show results on the application of the Generalized Hurst exponent tool to different financial time series and we will show the powerfulness of such tool to detect changes in markets behaviours, to differentiate markets accordingly to their degree of development, to asses risk and to provide a new tool for forecasting. We will also show an empirical relationship, to our knowledge the first on in the literature, which links a univariate property, i.e. the degree of multiscaling behaviour of a time series, to a multivariate one, i.e. the average correlation of the stock log-returns with the other stocks traded in the same market and discuss its implications.

C1434: **Selection and estimation of fractional and multifractional models**

Presenter: **Matthieu Garcin**, Leonard de Vinci Pole Universitaire, France

The Hurst exponent describes the scaling properties of a time series. One also often links its value to the persistence of the series and consequently to one's ability to forecast it: if $H = 1/2$ there is no autocorrelation, if $H > 1/2$ the series is persistent, and if $H < 1/2$ the series is anti-persistent. However, the interpretation of the Hurst exponent strongly depends on the model describing the dynamic. We are interested in three classes of models: the fractional Brownian motion (fBm), multifractional Brownian motions (mBm), and transforms of a fBm. The mBms are extensions of the fBm and rely on the assumption that the Hurst exponent is time-varying or even is a random process, whereas it is a constant for the fBm. Transforms of the fBm, such as the fractional Ornstein-Uhlenbeck process or the Lamperti transform of a fBm, are of practical interest, for example in the fixed income world, to model stationary processes. We expose the specificities of the estimation of the Hurst exponent for all these models as well as the way one can forecast such series, using accuracy metrics that are relevant in the perspective of a portfolio manager. We also address the issue of selecting the proper fractional or multifractional model, based on the data.

C1534: **Credit scoring: Moving beyond the traditional approach with random forests and artificial neural networks**

Presenter: **Samuel Stephan**, Universita Paris I Panthaon - Sorbonne, France

Co-authors: Matthieu Garcin

Traditional approaches, such as the logistic regression, are widespread in credit scoring because they are fast to implement, give quite accurate results, and allow for explainability. More advanced methods, such as ensembles (e.g. random forest, gradient boosting) or ANN (Artificial Neural Network), stay underused because of their complexity and their resulting lack of transparency. Practitioners are reluctant to use these techniques because they are known to have a high variance due to their large number of parameters which can make them unstable on new data. Indeed, these algorithms require knowledge on hyperparameters tuning in order to get workable results. We have compared three models: a logistic regression as a benchmark, a random forest, and an ANN. These algorithms have been fitted on a real credit dataset on individual loans sold from 2015 to 2018 by a French bank. The results show that random forest and ANN outperform the traditional logistic regression and would benefit to be used in financial applications.

CO721 Room MAL B35 ADVANCES IN FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C0277: **Hedge fund return predictability in the presence of model risk**

Presenter: **Nikolaos Voukelatos**, University of Kent, United Kingdom

Co-authors: Christos Argyropoulos, Ekaterini Panopoulou, Teng Zheng

Hedge funds implement elaborate investment strategies that include a variety of positions and assets. As a result, there is significant time variation in the set of risk factors and their respective loadings which in turn introduces severe model risk in any attempt to model and forecast the hedge fund returns. We investigate the statistical and economic value of incorporating heteroscedasticity, non-normality, time-varying parameters, model selection risk and parameter estimation risk jointly in hedge fund return forecasting and fund of funds construction. Parameter estimation risk is dealt with by a time-varying parameter structure, while model selection uncertainty is mitigated by model averaging or model selection. We adopt a dynamic model averaging approach along with the conventional Bayesian averaging technique. Our empirical results suggest that accounting for model risk can significantly improve the hedge fund returns forecasting accuracy and consequently the performance of the hypothetical fund of funds.

C0701: **Dynamic estimation of information asymmetry risk in trading**

Presenter: **Jaideep Oberoi**, University of Kent, United Kingdom

Co-authors: Jim Griffin, Samuel Oduro

A new dynamic indicator is proposed for the risk of informed trading inferred from traded prices, volume and bid-ask spread. Several widely used measures of information asymmetry rely on decomposing traded volume into expected and unexpected components. The approach exploits the non-linear relationship between bid-ask spreads and volume. Both volume and spreads have been shown to be (partly) driven by an underlying liquidity process. We model a stochastic latent information process in a state-space model that allows us to identify different types of volume changes, and their implications for informed trading. We use a Bayesian procedure with a Gibbs sampling algorithm to estimate the model on a sample of 10 stocks over a 10 year period. The results are consistent with the opposing effects of informed and uninformed volume on the spread. We also find that both informed and uninformed trading are significantly persistent over the sample period.

C1163: **Forecasting spot price in the UK natural gas market**

Presenter: **Chih-Yueh Huang**, University of Kent, United Kingdom

Co-authors: Ekaterini Panopoulou, Stella Hadjiantoni

Natural gas is one of the most widely used energy resources in the UK as it is the main energy commodity for domestic heating and power generation. The pricing of natural gas prices is crucial for generators, power and gas suppliers, manufacturing plants, real estate agency and other downstream users since the movements of gas prices affect their revenue or costs. Different popular forecasting models, such as forecast combination, penalised regression, stepwise regression and principal components regression, are employed to analyse and forecast UK gas spot price. The result shows that more advanced techniques do not improve the forecasting performance based on Mean squared errors (MSE) ratio and success ratio in the out-of-sample periods. Then, we compare the energy procurement costs of three purchasing strategies, namely spot only, futures only and mixing strategy. The empirical evidence shows that comparing the spot forecast and the front-month futures price helps energy users reduce purchasing costs.

C1198: **The effects of macroeconomic variables on the cross-section of mutual funds returns: A threshold approach**
Presenter: **Christos Argyropoulos**, Lancaster University, United Kingdom
Co-authors: Christos Argyropoulos, Bertrand Candelon, Jean-Baptiste Hasse, Ekaterini Panopoulou
The role of macroeconomic variables on the cross-section of mutual funds returns is investigated. Specifically, it associates macroeconomic variables with regimes, during which the Fama-French 3-factor model is stable. It thus allows testing for the stability of the asset pricing model over the macroeconomic regime. Contrary to previous studies, macroeconomic variables are not introduced as a factor, nor a benchmark but as reflecting macroeconomic conditions. We find that the linearity of the Fama-French 3-factor model is strongly rejected and that macroeconomic variables define regimes of stability for asset pricing models.

CO216 Room MAL B36 MACHINE LEARNING IN FINANCE Chair: Jozef Barunik

C0633: **Structural models for firm bankruptcy prediction**
Presenter: **Ludovico Rossi**, Colegio Universitario De Estudios Financieros - CUNEF, Spain
The role of the Merton distance-to-default model in forecasting corporate bankruptcies is investigated. Given the high number of bankruptcy predictors proposed in the literature, we use Logistic Lasso regressions to perform variable selection and Post Lasso Logit regressions to estimate conditional hazard models. In contrast with the existing literature, the Merton distance-to-default model consistently predicts bankruptcies in all time periods and industries. This result is robust to different model specifications. Moreover, we show that the Merton distance-to-default model is one of the most accurate variables to predict bankruptcies. Out-of-sample forecasts confirm that Post Lasso Logit, which includes Merton distance-to-default, produce more accurate bankruptcy predictions than previous models proposed in the literature.

C0876: **Dynamic density forecasting using machine learning**
Presenter: **Lubos Hanus**, UTIA AV CR, v.v.i, Czech Republic
Co-authors: Jozef Barunik
The use of machine learning techniques is proposed to describe and forecast the conditional probability distribution of asset returns. We redefine the problem of forecasting of conditional probabilities looking from a different perspective than traditional ordered binary choice models. Using deep learning methods, we offer a better description of asset returns distribution. The study on the most liquid U.S. stocks shows that predictive performance of machine learning methods is promising out-of-sample. We provide a comparison of machine learning methods to the unordered and order binary choice models used by the literature.

C1280: **Asset pricing with quantile machine learning**
Presenter: **Martin Hronec**, Faculty of Social Sciences, Charles University in Prague, Czech Republic
Co-authors: Jozef Barunik
A large scale empirical test is performed for an asset pricing model based on agents with quantile utility preferences instead of the standard expected utility. Using machine learning methods, we predict quantiles of individual stock returns obtaining the whole forecasted distributions. We document heterogeneity in models parameters across different quantiles. We show that forecasting all quantiles together, using multi-task deep learning is better than forecasting quantiles individually. The forecasting models allow us to construct portfolios based on the whole distribution instead of just a conditional mean. We show the economic value added of looking at the whole forecasted distribution by forming quantile-based long-short portfolios, as well as favourably forecasting value-at-risk.

C1511: **Climate risks and stock returns**
Presenter: **Eugenio Carnemolla**, University of Lausanne and Swiss Finance Institute, Switzerland
Co-authors: Giuseppe Vinci
A novel measure of exposure to climate risks is constructed by applying textual analysis to the firms' annual reports. The relation between weather sensitivity and word counts of the filings is estimated using a supervised machine learning algorithm. We validate our text-based measure of climate risk by using information on the firm's geographical footprint. We find that stocks of climate-sensitive firms significantly underperform stocks of climate-resilient firms, suggesting that investors underreact to climate change risks associated with natural disasters. Our results are stronger following months with attention-grabbing weather events and for geographically concentrated firms. A trading strategy exploiting weather sensitivity earns an annualized five-factor alpha of 6.24%.

CO739 Room Montague FACTOR MODELS Chair: Eric Renault

C1458: **Factor models for conditional asset pricing**
Presenter: **Paolo Zaffaroni**, Imperial College London, United Kingdom
A methodology is developed for inference on no-arbitrage conditional asset pricing models linear in latent risk factors, valid when the number of assets diverges but the time series dimension is fixed, possibly very small. We show that the no-arbitrage condition permits to identify the risk premia as the expectation of the latent risk factors. This result paves the way to an inferential procedure for the factors' risk premia and for the stochastic discount factor, spanned by the latent risk factors. In our set up naturally, almost every feature of the asset pricing model is allowed to be time-varying including loadings, idiosyncratic risk and the number of risk factors. Several Monte Carlo experiments corroborate our theoretical findings. An empirical application based on individual asset returns data demonstrates the power of the methodology, allowing us to tease out the empirical content of the time-variation stemming from asset pricing theory.

C1463: **High-dimensional functional factor models**
Presenter: **Shahin Tavakoli**, University of Warwick, United Kingdom
Co-authors: Gilles Nisol, Marc Hallin
Theoretical foundations are set up for high-dimensional approximate factor models for a panel of functional time series (FTS). We first establish a representation result stating that if the first r eigenvalues of the covariance operator of a cross-section of N FTS are unbounded as N diverges and if the $(r+1)$th one is bounded, then we can represent each FTS as a sum of a common component driven by r factors, common to (almost) all the series, and a weakly cross-correlated idiosyncratic component (all the eigenvalues of the idiosyncratic covariance operator are bounded as N diverges). The model and theory are developed in a general Hilbert space setting that allows for panels mixing functional and scalar time series. We then turn to the estimation of the factors, their loadings, and the common components. We derive consistency results in the asymptotic regime where the number N of series and the number T of time observations diverge, thus exemplifying the "blessing of dimensionality" that explains the success of factor models in the context of high-dimensional (scalar) time series. The results encompass the scalar case, for which they reproduce and extend, under weaker conditions, well-established results. We provide numerical illustrations and an empirical illustration on a dataset of intraday S&P100 and Eurostoxx 50 stock returns, along with their scalar overnight returns.

C0782: **Skill and value creation in the mutual fund industry**
Presenter: **Patrick Gagliardini**, University of Lugano, Switzerland
Co-authors: Laurent Barras, Olivier Scaillet
A simple, nonparametric approach is developed for estimating the entire distribution of skill. The approach avoids the challenge of correctly specifying the distribution, and accommodates the need to study both the investment and trading dimensions of skill. We estimate different skill

measures (first-dollar alpha, size coefficient, value added) at the individual fund level using a large unbalanced panel factor model and estimate by kernel smoothing the density, quantiles and moments of the skill distribution. We show how to adjust the estimates for the asymptotic bias induced by the Error-in-Variable (EIV) problem in a large-N-large-T setting. The results show that most funds are skilled at detecting profitable trades, but unskilled at overriding capacity constraints. Aggregating both skill dimensions, we find overwhelming evidence that mutual funds produce significant value added. In addition, the active industry is (i) not concentrated because few funds are skilled on all dimensions, (ii) close to optimally sized as funds internalize the impact of capacity constraints, and (iii) in a strong bargaining position vis-a-vis the investors.

CO388 Room Woburn EMPIRICAL MACRO AND FINANCE Chair: Alessia Paccagnini

C0916: **Does trilemma speak Chinese?**
Presenter: **Georgios Magkonis**, University of Portsmouth, United Kingdom

Based on the limitations imposed by the trilemma, the trade-offs faced by the Chinese economy is examined. Taking into account the role of accumulation of foreign reserves we examine how binding the constraints are for the Chinese monetary authorities. Additionally, using a panel VAR with dynamic and static interdependencies as well as cross-sectional heterogeneities, we examine the monetary spillovers from China to a series of Asian economies. In this way, we measure the degree to which the Chinese trilemma constraints are exported to other countries. Consistent with previous research, the empirical evidence suggests that China's trilemma configurations are unique as China manages to achieve exchange rate stability, along with moderate financial liberalization, without losing its monetary autonomy. Furthermore, there are no significant spillovers to regional economies. Overall, trilemma does speak Chinese, but only for a short period.

C1179: **On the effect of exchange rate on production**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland
Co-authors: Isabella Blengini

The relationship between exchange rate and production has often been analysed in both theoretical and empirical literature. The theory suggests a quite clear relationship between exchange rate and output behavior: an appreciation of the currency, should make domestic production more expensive in the eyes of its international trade partners. At the same time, domestic consumers would experience an increase in their purchasing power abroad. In other terms, it would be cheaper for them to import goods from abroad. These two combined effects would reduce domestic and foreign demand for domestically produced goods, and that would generate a contraction in production. Nevertheless, the empirical evidence does not confirm these theoretical predictions. We conduct a panel VAR analysis which includes OECD and developing countries and several industries to study how exchange rate movements and monetary policy uncertainty affect these economies sector by sector.

C1218: **Mind the gap: Stylized dynamic facts and structural models**
Presenter: **Filippo Ferroni**, Federal Reserve Bank of Chicago, United States
Co-authors: Fabio Canova

The purpose is to study what happens to identified shocks and to dynamic responses when the structural model features q disturbances and m endogenous variables, $q = m$, but only $m_1 < q$ variables are used in the empirical model. Aggregation create problems. Appropriate theoretical restrictions may be insufficient to obtain the structural disturbances and the dynamics they produce. Identified shocks do not necessarily combine structural disturbances of the same type. Instead, they are linear combinations of current and past values of all structural disturbances. The theory used to interpret the data and the disturbances it features determines whether an empirical model is too small or not. An example highlights the magnitude of the distortions and the steps needed to reduce them. We revisit previous evidence regarding the transmission of house price shocks.

C1260: **Bank capital and credit supply shock in Ireland: A narrative approach**
Presenter: **Fabio Parla**, Central Bank of Ireland, Ireland
Co-authors: Martin Obrien, Sofia Velasco, Maria Woods, Michael Wosser

Structural Vector Autoregression analysis is suggested to be used in order to assess the effects of changes in bank capital in Ireland, over the 1998-2017 time span. In particular, the focus is on estimating to what extent an exogenous disturbance to banks' capital impacts on a set of macro-financial variables. The endogenous variables considered are proxies of economic activity, property price, credit aggregates and bank capital, observed at quarterly frequency. The positive structural exogenous shock hitting the VAR is identified by combining theory-driven sign restrictions and narrative information. The empirical findings suggest that a positive shock to bank's capital, interpreted as a negative credit supply shock, has a non-negligible impact on the macroeconomic variables. We find significant impacts on variables such as credit, interest rate and house prices in response to the exogenous shock to capital.

CO759 Room Chancellor's Hall NONPARAMETRIC/SEMIPARAMETRIC ESTIMATION AND TESTING Chair: Luke Nicholas Taylor

C0290: **Functional sequential treatment allocation**
Presenter: **Bezirgen Veliyev**, Aarhus University, Denmark
Co-authors: Anders Kock, David Preinerstorfer

A treatment allocation problem with multiple treatments is studied, in which the individuals to be treated arrive sequentially. The quality of a treatment is allowed to be measured through a general (combination of) functionals of the underlying distribution of treatment outcomes, including inequality-, welfare- and poverty-measures. The goal of the policy maker is to minimize maximal expected regret compared to always assigning the unknown best treatment to all individuals. We first show that a natural approach to the policy maker's problem based on conducting an RCT to learn which treatment is best can incur very high maximal expected regret irrespective of the decision rule employed following the RCT. Motivated by this finding we study the Functional Upper Confidence Bound (FUCB) policy, which interweaves exploration and exploitation, and show that it performs better than any two-step policy based on an RCT. Furthermore, we show that, irrespective of the functional of interest, the expected regret incurred by the FUCB policy is near-minimax optimal. Next, we study the case of heterogeneous treatment outcome distributions by introducing covariates and show that the FUCB policy is minimax optimal over a broad class of treatment outcome distributions under minimal assumptions.

C0467: **Minimax learning for average regression functionals with an application to electoral accountability and corruption**
Presenter: **Chen Qiu**, London School of Economics, United Kingdom

A new minimax methodology is proposed to estimate average regression functionals, which cover many empirical problems including average treatment effect. Featured in penalized series space, this strategy exploits minimax property of a vital nonparametric component of average regression functional and aims to directly control key remainder bias. We then construct a new class of estimators, called minimax learners, and study their asymptotic properties when number of controls over sample size goes to zero, constant and infinity, respectively. Root-n normality is established under weak conditions for all three cases. Minimax learners are fast to implement due to their minimum distance representation. In simulations where selection bias is mild, they behave more stably, show less mean square error and do not over control. When applied to a previous work that studies effect of electoral accountability on corruption, minimax learners behave less erratically and lead to more coherent conclusion, even when number of controls becomes very large.

C0894: **Identification of type I and type II error probabilities: Judging the justice system**
Presenter: **Luke Nicholas Taylor**, Aarhus University, Denmark
Co-authors: Shin Kanaya

New nonparametric identification results for the misclassified binary choice model are provided, and it is shown that the misclassification probabilities can be interpreted as type I and type II error probabilities in a range of important empirical settings. We use our strategy to answer: what is the likelihood that a mistake is made in a courtroom? Using US case-level data, we estimate both the probability of convicting an innocent defendant and the probability of acquitting a guilty defendant. Furthermore, we allow these estimates to depend on case and defendant characteristics in a nonparametric way.

C1059: **M-Estimation with isotonic estimator plugged-in**
Presenter: **Mengshan Xu**, London School of Economics and Political Science, United Kingdom

A semiparametric estimator is studied, where the moment condition associated with it contains a nuisance monotone function, which is estimated nonparametrically by isotonic regression. We show that the properties of isotonic regression satisfy s previous framework, and we obtain a semiparametric estimator which is root-n consistent and asymptotically normally distributed. We give the conditions that this estimator reaches semiparametric variance bound. We show this structure can help practitioners to obtain tuning-parameter-free estimators for many semiparametric models, including partial linear model, single-index model and particularly, average treatment effect model.

CC818 Room MAL 352 CONTRIBUTIONS IN RISK ANALYSIS Chair: Raimund Kovacevic

C0348: **Deep time-series feature extraction in credit scoring models**
Presenter: **Jui-Yu Lin**, National Chiao Tung University, Taiwan
Co-authors: Huei-Wen Teng, Yu-Huai Yu, Kai-Shiang Fan, Yi-Chia Lin

Credit scoring models predict the default of a credit card or loan holder and are of considerable importance in the banking system. We investigate the use of time-series extracted feature using deep learning neural network for predicting default in credit scoring models. With borrowers' payment history records, we extract time-series patterns and predict default risk by using Convolutional Neural Networks and Recurrent Neural Networks. We compare a set of machine learning methods with and without our deep extracted time-series features using two data sets, one is from an open source and the other is form a major bank in Taiwan. Our numerical results show that the performance can be improved with time-series extracted feature in terms of AUC.

C0215: **Controlling tail risk measures with estimation error**
Presenter: **Tetsuya Kaji**, University of Chicago, United States
Co-authors: Hyungjune Kang

Assessment of risk and its control play an important role in investment decision making, financial regulations, actuary science, and operations research. In practice, accuracy of estimated risk is subject to estimation error. While the estimation error can be estimated in many cases, it remains a question as to how the error thus estimated can be incorporated into actual control of the true but unobservable risk. We propose the class of risk measures, called the tail risk measures, that give the upper bounds below which the quantities of interest fall with probability at least as much as a pre-specified confidence level. We show that a simple rule based on the Bonferroni inequality can control a tail risk measure at a desired level, even when the true risk is unknown and needs to be estimated. Most popular risk measures such as Value-at-Risk and expected shortfall are interpreted as tail risk measures. For coherent tail risk measures, the true risk of any combination of assets can be controlled by knowledge of estimated risk and estimated error of individual assets. Empirical applications illustrate how the proposed concept can be applied to practical risk control problems.

C1968: **Bayesian estimation of realized EGARCH model to forecast tail risks**
Presenter: **Vica Tendenan**, The University of Sydney, Australia
Co-authors: Richard Gerlach, Chao Wang

A Bayesian framework is developed for the realized exponential generalized autoregressive conditional heteroskedasticity (Realized EGARCH) model that uses multiple realized volatility measures for the modelling of a return series. The Realized EGARCH model is extended by adopting not only a Gaussian distribution for the return equation, but also a standardized student-t and a skewed-t distribution. Meanwhile, a Gaussian distribution is adopted for the innovations in the measurement equation(s). Different types of realized measures are considered, such as realized variance, realized kernel, and realized range. The Bayesian estimation is conducted by employing Markov chain Monte Carlo (MCMC) procedures by using the robust adaptive Metropolis algorithm (RAM) in the burn in period and the standard random walk Metropolis in the sample period. The Bayesian estimators are compared with maximum likelihood estimators and show more favourable results. We apply the model to six international equity index markets and forecast tail risks such as value at risk (VaR) and expected shortfall (ES). The one-step-ahead forecast of the tail risks is conducted for over a period of 1000 days. The forecast performance of the model is evaluated via VaR and ES backtests.

C1773: **Tail risks, asset prices, and investment horizons**
Presenter: **Matej Nevrla**, Czech Academy of Sciences, Czech Republic
Co-authors: Jozef Barunik

The aim is to examine how extreme market risks are priced in the cross-section of asset returns at various horizons. Based on the frequency decomposition of covariance between indicator functions, we define the quantile cross-spectral beta of an asset capturing tail-specific as well as horizon-, or frequency-specific risks. Further, we work with two notions of frequency-specific extreme market risks. First, we define tail market risk that captures dependence between the extremely low market as well as asset returns. Second, extreme market volatility risk is characterized by dependence between extremely high increments of market volatility and extremely low asset return. Empirical findings based on the datasets with long enough history, 30 Fama-French Industry portfolios, and 25 Fama-French portfolios sorted on size and book-to-market support our intuition. We reach the same conclusion using stock-level data as well. These results suggest that both frequency-specific tail market risk and extreme volatility risks are significantly priced and our five-factor model provides an improvement over specifications considered by previous literature.

CC826 Room MAL 353 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS Chair: Yiannis Karavias

C1603: **A stochastic volatility model with two macro-financial components**
Presenter: **Yuze Liu**, University of Cologne, Germany

The use of macroeconomic variables to improve the accuracy of daily financial volatility estimation has been discussed in the past years. Volatility is typically decomposed into two components: a short-run component from the classical financial volatility estimation, and a long-run component corresponding to macroeconomic determinants. However, there are some important limitations: First, high-frequency volatility is usually estimated by using the GARCH-framework. Second, the low-frequency macroeconomic variables are included via discrete piecewise constant step-functions during a month or a quarter. A new and simple stochastic volatility model with two components is proposed. The short- and long-run component are mapped into two latent variables, where the long-run component is varying simultaneously with the short-run component at the high-frequency. Despite the increased flexibility of the newly proposed model, it retains the classical state-space representation which can be efficiently estimated by using standard Kalman Filter techniques. The finite sample properties are investigated by using extensive Monte Carlo simulations. The proposed model is compared against the classical one component SV and GARCH models, as well as two components GARCH-MIDAS models. As an empirical application, the US stock market volatility is investigated.

C1634: **When does attention matter: The effect of investors' attention on stock market volatility during news releases**
Presenter: **Daniele Ballinari**, University of St Gallen, Switzerland

Co-authors: Francesco Audrino, Fabio Sigrist

Empirical and theoretical studies have shown that measures of investor attention have a positive impact on future stock market volatility and trading volume. We address an often overlooked question: Is investor attention always relevant or is its effect on volatility varying depending on the release of new information? Constructing attention measures from two online social media platforms for 360 large cap US stocks in the S&P 500, we analyse the impact of investors' attention on future volatility in a fixed-effect panel framework. The results show that attention measures are more informative for next day's volatility when both scheduled and unscheduled news articles are released. The impact of investors' attention is even larger when articles are published after trading hours or unexpected news is released. In particular we find the largest effect during the release of unscheduled news articles about a company's fundamentals. These findings are confirmed by out-of-sample prediction results and an economic application.

C0694: **Encompassing tests for higher-order elicitable functionals**

Presenter: **Julie Schnaitmann**, Universitat Konstanz, Germany

Co-authors: Timo Dimitriadis

Encompassing tests are introduced for forecasts of higher-order elicitable functionals such as the variance and the Expected Shortfall (ES). Encompassing tests rely on the existence of strictly consistent loss functions for the forecasted functionals under consideration, which do not exist for the variance and the ES. However, for these functionals, such loss functions exist for the pairs (mean, variance) and (quantile, ES). We utilize these joint loss functions in order to introduce joint encompassing tests for the quantile and the ES as well as stand-alone encompassing tests for the ES. These tests provide a theoretical justification for forecast combination of the ES when encompassing is rejected. We show through simulation studies that all the proposed tests are reasonably sized and exhibit good power properties against general alternatives in typical financial applications. In the empirical application, we apply encompassing tests in order to demonstrate the superiority of forecast combination methods for the ES.

C0197: **Periodic dynamic conditional correlations in bond markets**

Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus

Co-authors: Nektarios Michail, Demetris Koursaros

The dynamic conditional correlation model of Engle is extended building on a previous extension to allow periodic day specific conditional correlations of shocks across international bond markets. When applied to the intra-week interactions among bond markets around the globe, over the period from 1999 2019, we find very strong evidence of periodic conditional correlations for the shocks.

CC817 Room MAL 354 CONTRIBUTIONS IN BAYESIAN ECONOMETRICS Chair: Tore Kleppe

C0184: **Investigating the role of money in the identification of monetary policy behavior: A Bayesian DSGE perspective**

Presenter: **Qing Liu**, Tsinghua University, China

The aim is to estimate an enriched version of the mainstream medium-scale DSGE model which features non-separability between consumption and real money balance in household's utility and a systematic response of the policy rate to money growth. The estimation results show that money is a significant factor in the monetary policy rule, without which it may lead to biased estimates of the model. In contrast to earlier studies that rely on small-scale models, the merits of using a sufficiently rich model is stressed. First, it delivers different results, such as the role of non-separability between consumption and real money balance in preference. Second, the rich dynamics embedded in the model allows us to explore the responses of a larger set of macroeconomic variables, and thus such a model is more informative on the effects of the shocks. Third and also most importantly, it avoids the possible pitfalls of small-scale model, which assures more reliable inferences on the role of money over business cycles.

C1677: **Constrained Bayesian SVARs: Two specifications in the world crude oil market modelling**

Presenter: **Yunyi Zhang**, The University of Warwick, United Kingdom

Constrained estimation of Bayesian structural vector autoregressions is proposed. We call it C-BSVARs. C-BSVARs numerically determine a constrained solution for parameters of interests, and then utilises a random-walk Metropolis–Hastings algorithm to sample the posteriors. We illustrate that C-BSVARs enable a flexible and efficient platform for identifications via two specifications in the world crude oil market. The first specification uses exact identifications, and finds their economic findings to be coherent. However, the estimations of oil demand elasticities are too diffuse within the truncated range, and are sensitive to the choice of seed for random numbers. The magnitude of oil demand elasticities is of concern, since it decides the existence of oil price endogeneity and relative importance of oil demand and supply shocks on influencing oil market fluctuations. The second specification, therefore, proposes a lower-bound uncertainty on the short-run oil demand elasticity for use, which is randomly sampled from a truncated Student t distribution. Adding the additional restriction, C-BSVARs sharpen the inference of impulse response functions and shift the oil demand elasticities towards zero. Further, the restriction of uncertainties for elasticities provides a novel way for identifying key structural parameters that are nonlinear.

C1513: **Generalized Poisson difference autoregressive processes**

Presenter: **Giulia Carallo**, Ca' Foscari University of Venice, Italy

Co-authors: Roberto Casarin, Christian Robert

In many real-world applications, time series of counts are commonly observed given the discrete nature of the variables of interest. A new stochastic process with values in the set Z of integers with sign is introduced. The increments of the process are generalized Poisson differences and the dynamics has an autoregressive structure. In order to deal with the time-varying nature of the parameters, we introduce an integer-valued GARCH process. We study the properties of the process and exploit the thinning representation to derive stationarity conditions and distribution of the process. We develop a Bayesian inference and an efficient posterior approximation procedure based on Markov chain Monte Carlo, that allow us to make more tractable the likelihood function of the GPD and to include in the estimation prior information about the parameters. Numerical illustrations on both simulated and real data show the effectiveness of the proposed inference.

C1972: **Investigating the contagion effect with using Bayesian Copula-GARCH models**

Presenter: **Justyna Mokrzycka**, Cracow University of Economics, Poland

The aim is to present a concept of identifying the contagion effect consisting in comparing a posteriori probabilities of two Bayesian dynamic Copula-GARCH models with t Student or normal copula. In case of a higher a posteriori probability of model with modified dynamics and a positive value of an additional parameter, these results are proposed to be interpreted as a confirmation of contagion effect on the financial markets. The contagion effect is understood in a narrow sense, as an increase in linkages between financial markets. Results for simulated and empirical data are presented. Using Copula-GARCH models enables a determination of asymmetric structures of dependencies between financial markets, and moreover, in separation from a dynamics of volatility in individual markets. Student's conditional marginal distributions may have a different number of degrees of freedom, the dependence may be asymmetric, i.e. coefficients of upper and lower tail dependencies may have different values. A formal choice of a proper structure of dependence in Copula-GARCH models is possible by applying Bayesian inference.

CG445 Room MAL 351 CONTRIBUTIONS IN PORTFOLIO OPTIMIZATION II Chair: Tomer Shushi

C1953: **Tail risks in vast portfolio selection: A comparison of penalized quantile versus expectile models**

Presenter: **Rosella Giacometti**, University of Bergamo, Italy

Co-authors: Gabriele Torri, Sandra Paterlini

Estimating in an accurate way, and optimally controlling tail risk, is of utmost importance for building portfolios with desirable properties for investors, especially in presence of a large set of assets. In recent years, the financial literature witnessed an increase of interest towards expectiles as an alternative to more common risk measures such as Value at Risk (VaR) and Expected Shortfall (ES). Such a family of measures has good theoretical properties (it is the only risk measure that is both coherent and elicitable), and has a relevant financial interpretation (it can be thought as the amount of money that should be added to a position in order to have a sufficiently high gain-loss ratio). We combine expectile and quantile approaches with regularization to build optimal portfolio models, with the aim of providing parsimonious and robust portfolios with better out-of-sample performances. Simulation and real-world analysis allow us to critically discuss pros and cons of the proposed methods when compared to state-of-art benchmarks.

C1954: **Penalized expectiles optimal portfolios**
Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Rosella Giacometti

Expectiles are risk measures increasingly popular in recent years among academics and practitioners, thanks to their good theoretical properties: they are the only risk measure that is both coherent and elicitable. Moreover, they have an intuitive economic explanation and interesting algebraic connections can be established with Value at Risk (VaR) and Expected Shortfall (ES). Recent works explored their usage in portfolio optimization, showing how to build optimal risk-return portfolios using expectiles as risk measure. However, real-world application to portfolios with a large number of assets are limited by estimation error, that typically leads to bad out of sample performances. We propose a novel derivation of the linear programming formulation of the minimum EVaR portfolio, similar to the ones available in the literature, but computationally faster and characterized by a straightforward economic interpretation. We also introduce a ridge penalization to the portfolio weights in order to improve the finite sample performances, and we test the model on a variety of datasets.

C0564: **Dynamic modeling of the global minimum variance portfolio weights**
Presenter: **Laura Reh**, University of Cologne, Germany
Co-authors: Fabian Krueger, Roman Liesenfeld

A novel dynamic approach is proposed to forecast the weights of the global minimum variance portfolio (GMVP). We exploit the fact that the GMVP weights can be obtained as the population coefficients of a linear regression of one benchmark return on a vector of return differences. This enables us to derive a consistent loss function from which we can infer the optimal GMVP weights without imposing any distributional assumptions on the returns. In order to capture time variation in the assets' conditional covariance structure, we model the portfolio weights through a Recursive Least Squares scheme as well as by Generalized Autoregressive Score type dynamics. Sparse parameterizations ensure scalability with respect to the number of assets. An empirical analysis of daily and monthly financial returns shows that the model performs well in-and out-of-sample in comparison to existing approaches.

C0190: **Multi-asset investing: Correlation structure between Canadian real estate and equity markets**
Presenter: **Ivan Medovikov**, Brock University, Canada
Co-authors: Jean-Francois Lamarche

Investors often view real estate as means of diversifying market risk away from equity portfolios. We put this notion to the test in the Canadian context and study nature of dependence between returns to an index of Canadian real estate trusts (REITs) and public equities. Using a semi-parametric copula model we show dependence between the two asset classes to be non-linear and skewed toward distribution tails. In particular, we find that linkages between real estate and equities to be strongest and also positive during market downturns, suggesting that real estate may be poor choice when it comes to protecting against stock market shocks.

CG443 Room MAL 355 CONTRIBUTIONS IN TIME SERIES I Chair: Richard Luger

E1823: **A robust procedure to build dynamic factor models with cluster structure**
Presenter: **Pedro Galeano**, Universidad Carlos III de Madrid, Spain
Co-authors: Andres M Alonso, Daniel Pena

Dynamic factor models provide a useful way to model large sets of time series. These data often have heterogeneity and cluster structure and the formulation and estimation of dynamic factor models should be adapted to these features. A procedure is presented to fit Dynamic Factor Models with Cluster Structure (DFMCS), where some of the factors are global and others group-specific, to heterogeneous data that may include multivariate additive outliers and level shifts. The procedure starts with an initial cleaning of the times series from outlying effects. Then, a first estimation of the possible factors is applied to the cleaned data and these factors are used to build the common component of each series. The groups are found by studying the joint dependency of these common components. Then additional factors are estimated by using the series in each cluster and, finally, all the factors found are classified as global or group-specific. We show in a Monte Carlo study that the procedure works well and seems to be better than other alternatives in terms of estimation of factors and loadings as well as in terms of misclassification rates for the series. An example of an electricity market is presented to illustrate the advantages of cleaning for outliers and taking into account the cluster structure for understanding and forecasting.

E0359: **A model-averaging approach for functional-coefficient regression**
Presenter: **Yuying Sun**, Academy of Mathematics and System Science, Chinese Academy of Sciences, China
Co-authors: Zongwu Cai, Shouyang Wang

Model averaging aims at providing an insurance against selecting a poor forecast model. All existing model averaging approaches in the literature are designed with constant combination weights. Little attention has been paid to functional weighing in model averaging, which is more realistic in economics and finance. A novel model averaging estimator is proposed which selects optimal functional combination weights by minimizing a local leave-subject-out cross-validation criterion. It is shown that the proposed functional leave-subject-out cross-validation model averaging (FLsoMA) estimator is asymptotically optimal in the sense of achieving the lowest possible local squared error loss in a class of functional model averaging estimators. Under a set of regularity assumptions, the FLsoMA estimator is root-Th consistent. A simulation study and an empirical application highlight the merits of the proposed FLsoMA estimator relative to a variety of popular estimators with constant model averaging weights and model selection.

E1919: **Estimation of the long-run error variance in nonparametric regression with time series errors**
Presenter: **Marina Khismatullina**, University of Bonn, Germany
Co-authors: Michael Vogt

A new difference-based estimator of the long-run error variance for nonparametric regression is proposed in the case that the error terms have an autoregressive structure. Such an estimator is required for virtually all inferential procedures in the context of nonparametric regression. Our proposed estimator improves on existing methods in several respects. First, the estimator produces accurate estimation results even when the AR process is quite persistent. Second, it produces accurate results even in the presence of a very pronounced regression function. These properties are illustrated by a simulation study that compares the proposed estimator with existing ones.

E1834: **The mHMMbayes R package for fitting mixed hidden Markov models using Bayesian estimation**

Presenter: **Emmeke Aarts**, Utrecht University, Netherlands

The mixed hidden Markov model (HMM) is a generalization of the well-known hidden Markov model, tailored to accommodate (intense) sequential data of multiple individuals or objects simultaneously. Using a mixed (also known as multilevel) framework, we allow for heterogeneity in the model parameters (transition probability matrix and conditional distribution), while estimating one overall HMM. The package mHMMbayes a useful tool to estimate mixed HMMs in the programming language R, and the only CRAN package that allows fitting such and fixed Bayesian HMM models. The model has a great potential of application in many fields, such as the social sciences and medicine. The model can be fitted on multivariate data with a categorical distribution, and include individual level covariates (allowing for e.g., group comparisons on model parameters). Parameters are estimated using Bayesian estimation utilizing the forward-backward recursion within a hybrid Metropolis within Gibbs sampler. The package also includes various automated visualizations of the fitted model, a function to simulate data, and a function to obtain the most likely hidden state sequence for each individual using the Viterbi algorithm.

Monday 16.12.2019 10:50 - 12:55 Parallel Session N – CFE-CMStatistics

EI010 Room Beveridge Hall SENSITIVITY ANALYSIS FOR UNCHECKABLE ASSUMPTIONS Chair: Michael Daniels

E0160: **Bayesian parametric approach to handle missing longitudinal outcome data in trial-based health economic evaluations**
Presenter: **Gianluca Baio**, University College London, United Kingdom
Co-authors: Michael Daniels, Andrea Gabrio

Trial-based economic evaluations are typically performed on cross-sectional variables, derived from the responses for only the completers in the study, using methods that ignore the complexities of utility and cost data (e.g. skewness and spikes). We present an alternative and a more efficient Bayesian parametric approach to handle missing longitudinal outcomes in economic evaluations, while accounting for the complexities of the data. We specify a flexible parametric model for the observed data and partially identify the distribution of the missing data with partial identifying restrictions and sensitivity parameters. We explore alternative non-ignorable scenarios through different priors for the sensitivity parameters, calibrated on the observed data. Our approach is motivated by, and applied to, data from a trial assessing the cost-effectiveness of a new treatment for intellectual disability and challenging behaviour.

E0161: **Inference taking into account sampling variation and uncertainty due to (un)testable model assumptions**
Presenter: **Xavier de Luna**, Umea University, Sweden

With very large datasets statistical inference cannot treat modelling assumptions as if they were given. While sampling variation may be large when small samples are analysed, with very large samples this is not anymore the case and it becomes important to acknowledge uncertainty due to modelling assumptions, since the latter may actually dominate sampling uncertainty. We advocate the use of confidence intervals taking into account both sampling variation and uncertainty due to modelling assumptions. We distinguish two types of modelling assumptions: those that can be investigated empirically with the data at hand and those that cannot. If model assumptions are tested, this should be taken into account in the final inference. Model assumptions which are not testable may be encompassed in a collection of a priori reasonable scenarios, implying a recognition that the parameter of interest may not be point identified but that only subsets of the parameter space, e.g. an identification interval, can be retrieved asymptotically. We review existing results and propose a general strategy for obtaining confidence intervals with desired coverage in the presence of testable and untestable model assumptions. We give illustrations by considering several examples where assumptions on missing data mechanisms are made in the context of classical parametric regression, as well as flexible semi-parametric estimation of causal parameters.

E0162: **Simulation-based sensitivity analysis for interference in observational studies with unmeasured links**
Presenter: **Fabrizia Mealli**, University of Florence, Italy
Co-authors: Alessandra Mattei, Laura Forastiere

In causal studies where the commonly invoked "no-interference" assumption is arguable, ignoring interference may lead to very misleading inferences. In observational studies, information on links between units is usually unavailable and interference cannot be taken into account. In a way, the neighborhood treatment can be seen as an unmeasured confounder. We propose to face this issue by developing a Bayesian simulation-based sensitivity analysis to the violation of the no-interference assumption, where we repeatedly i) draw a set of sensitivity parameters from a prior distribution, ii) simulate potential confounders, and iii) reestimate the posterior distribution of the effect of interest after adjusting for the simulated confounders. We propose a model to generate the unmeasured links, which carries our belief on the level of interference and on the level of association between the individual and the neighborhood treatments. If we assume interference to operate only through a function of the vector of neighbors treatments, after a network is drawn we can compute such function and estimate the direct effect of the treatment taking interference into account. Different functions can be used. This approach has the additional advantage of adjusting for neighborhood and network covariates

EO356 Room CLO B01 RECENT DEVELOPMENTS IN FUNCTIONAL DATA ANALYSIS Chair: Ping-Shou Zhong

E0451: **Functional single index quantile regression models**
Presenter: **Peijun Sang**, University of Waterloo, Canada
Co-authors: Jiguo Cao

It is known that functional single index regression models can achieve better prediction accuracy than functional linear models or fully nonparametric models, when the target is to predict a scalar response using a function-valued covariate. However, the performance of these models may be adversely affected by extremely large values or skewness in the response. In addition, they are not able to offer a full picture of the conditional distribution of the response. Motivated by using trajectories of PM10 concentrations of last day to predict the maximum PM10 concentration of the current day, a functional single-index quantile regression model is proposed to address those issues. A generalized profiling method is employed to estimate the model. Simulation studies are conducted to investigate the finite sample performance of the proposed estimator. We apply the proposed framework to predict the maximal value of PM10 concentrations based on the intraday PM10 concentrations of the previous day.

E0457: **Structure identification and sparse learning for image-on-scalar regression with application to imaging genetics studies**
Presenter: **Xinyi Li**, University of North Carolina at Chapel Hill, United States
Co-authors: Lily Wang, Huixia Judy Wang

High-dimensional image-on-scalar regression is considered, where the spatial heterogeneity of covariate effects on imaging responses is investigated via a flexible partially linear spatially varying coefficient model. To tackle the challenges of spatial smoothing over the imaging responses complex domain consisting of regions of interest, we approximate the spatially varying coefficient functions via bivariate spline functions over triangulation. We first study estimation when the active constant coefficients and varying coefficient functions are known in advance. We then further develop a unified approach for simultaneous sparse learning and model structure identification in the presence of ultra-high-dimensional covariates. Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal for constant coefficient estimators. The method is evaluated by Monte Carlo simulation studies and applied to a dataset provided by the Alzheimers Disease Neuroimaging Initiative.

E0476: **Adaptive lasso for the Cox regression with interval censored and possibly left truncated data**
Presenter: **Chenxi Li**, Michigan State University, United States
Co-authors: Daewoo Pak, David Todem

A penalized variable selection method is proposed for the Cox proportional hazards model with interval censored data. A penalized nonparametric maximum likelihood estimation is conducted with an adaptive lasso penalty, which can be implemented through a penalized EM algorithm. The method is proven to enjoy the desirable oracle property. We also extend the method to left truncated and interval censored data. Our simulation studies show that the method possesses the oracle property in samples of modest sizes and outperforms available existing approaches in many of the operating characteristics. An application to a dental caries data set illustrates the method's utility.

E1191: **Coherent mortality forecasting by weighted multilevel functional principal component approach**
Presenter: **Bo Wang**, University of Leicester, United Kingdom
Co-authors: Ruhao Wu

In human mortality modelling, if a population consists of several subpopulations it is desirable to model their mortality rates simultaneously while

taking into account the heterogeneity among them. Under closely related social, economic and biological backgrounds, mortality patterns of these subpopulations are expected to be non-divergent in the future. In this work we propose a weighted multilevel functional principal component analysis method for coherent mortality modelling, in the sense that the life expectancy in different populations does not diverge in the long run. We treat the mortality rates of subpopulations within a large population as a set of multilevel functional data and use the weighted multilevel functional principal component analysis to extract core information from the functional data and analyse them at multilevel scale, so that the model incorporates both overall information from the population as a whole and specific information from the subpopulations. The proposed model is applied to sex-specific data for nine developed countries, and the results show that the model outperforms some existing models developed in the literature in terms of overall forecasting accuracy.

E1575: **Order-restricted inference for means with missing values**
Presenter: **Heng Wang**, University of Illinois at Chicago, United States

Missing values appear very often in many applications, but the problem of missing values has not received much attention in testing order-restricted alternatives. Under the missing at random (MAR) assumption, we impute the missing values nonparametrically using kernel regression. For data with imputation, the classical likelihood ratio test designed for testing the order-restricted means is no longer applicable since the likelihood does not exist. A novel method is proposed for constructing test statistics for assessing means with an increasing order or a decreasing order based on jackknife empirical likelihood (JEL) ratio. It is shown that the JEL ratio statistic evaluated under the null hypothesis converges to a chi-bar-square distribution, whose weights depend on missing probabilities and nonparametric imputation. Simulation study shows that the proposed test performs well under various missing scenarios and is robust for normally and nonnormally distributed data. The proposed method is applied to an Alzheimer's disease neuroimaging initiative data set for finding a biomarker for the diagnosis of the Alzheimer's disease.

E2012: **A multiplier bootstrap approach to one-way ANOVA for functional data**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Miles Lopes, Hans-Georg Mueller

A new approach to the problem of one-way ANOVA for functional data is proposed based on basis expansion and our recent development of multiplier bootstrap for high-dimensional data with weak variance decay. We show that, the test not only admits the root-n consistency, but also has a size of a nearly parametric rate. Numerically, we demonstrate that it is comparable to existing tests in general situations, while exhibits a clear advantage when the problem is hard in the sense that the functional data are rough and/or the signals are weak.

EO486 Room MAL B02 ADVANCED STATISTICAL MODELLING FOR BIOMEDICAL DATA Chair: Andreas Mayr

E0328: **The effect of data aggregation on dispersion estimates in count data models**
Presenter: **Jochen Einbeck**, Durham University, United Kingdom
Co-authors: Adam Errington, David Endesfelder, Jonathan Cumming

For the modelling of count data, it is frequently convenient to aggregate the raw data over certain subgroups. Under the Poisson law, count data can be aggregated (over subgroups with the same predictor configuration) without information loss since the mean is the sufficient statistics for the Poisson parameter. This result remains true if the Poisson assumption is relaxed towards Quasi-Poisson, where one can also show that, assuming conditional independence of the raw counts, the dispersion of the aggregated data is the same as that of the raw counts. However, at this stage, problems start to creep in, which appear to have been largely overlooked in the literature. Firstly, it turns out that the variance of dispersion estimates can increase considerably following aggregation. Secondly, and more importantly, one can show through theory and simulation that relatively small deviations from the independence assumption in the raw data (say, the presence of strings of correlated observations) can lead to dramatically shifted dispersion values after aggregation. Notably, what is affected here is the dispersion itself, not just its estimate! The phenomena are illustrated through count-valued biomarkers (dicentric chromosomes, DNA repair proteins) as used in radiation biodosimetry for the calibration of dose-response curves.

E0493: **An adaptive lasso Cox frailty model for time-varying covariates based on the full likelihood**
Presenter: **Andreas Groll**, Technical University Dortmund, Germany
Co-authors: Maike Hohberg

A method is proposed to regularize Cox frailty models that accommodates time-varying covariates and is based on the full likelihood. A particular advantage of this framework is the explicit modeling of the baseline hazard in a nonlinear way, e.g. via P-splines. Linear covariate effects are penalized using the lasso penalty. The estimation is based on a Newton-Raphson algorithm and makes use of local quadratic approximations of the penalty terms. Additionally, adaptive weights are included to stabilize the estimation. The full likelihood model can easily be extended by a wide class of frailty distributions including random intercepts and random slopes. The method is implemented in R in the function coxlasso and will be compared to other packages for regularized Cox regression in both simulation scenarios and a real data application.

E1012: **Flexible nonparametric Bayesian density regression via dependent Dirichlet process mixture models and penalised splines**
Presenter: **Maria Xose Rodriguez-Alvarez**, BCAM, Basque Center for Applied Mathematics, Spain
Co-authors: Vanda Inacio, Nadja Klein

In many real-life applications, it is of interest to study how the distribution of a (continuous) response variable changes with covariates. Dependent Dirichlet process (DDP) mixture of normals models, a Bayesian nonparametric method, successfully addresses such a goal. The approach of considering covariate independent mixture weights, also known as the single weights dependent Dirichlet process mixture model, is very popular due to its computational convenience, but can have limited flexibility in practice. To overcome the lack of flexibility, but retaining the computational tractability, a single weights DDP mixture of normals model is developed, where the component's means are modelled using Bayesian penalised splines (P-splines). We coin our approach as DDPps. A practically important feature of DDPps models is that all parameters have conjugate full conditional distributions, leading to straightforward Gibbs sampling. In addition, they allow the effect associated with each covariate to be learned automatically from the data. The validity of our approach is supported by simulations and applied to a study concerning the association of a toxic metabolite on preterm birth.

E0510: **Boosting with random selection of weak learners for variable selection in high-dimensional biomedical data**
Presenter: **Christian Staerk**, University of Bonn, Germany
Co-authors: Andreas Mayr

Statistical boosting is a promising alternative to popular regularization methods such as the Lasso for modelling high-dimensional biomedical data with many possible explanatory variables: early stopping of the algorithm leads to implicit regularization and variable selection, enhancing the interpretability of the final models. Traditionally, the class of possible weak learners is fixed for all iterations of Boosting and consists of simple learners including only one explanatory variable at a time. Furthermore, the choice of the number of Boosting iterations is typically guided by optimizing the predictive performance of the resulting models, leading to models which often include unnecessarily large numbers of noise variables. We propose modifications of L_2Boost for variable selection in high-dimensional models which aim at addressing the potential issues described above. The modifications are based on an adaptive random selection of different classes of weak learners in each Boosting iteration. The considered classes include weak learners with several variables so that multiple coefficients can be updated at a single iteration. Furthermore, the

proposed modifications of L_2Boost can impose an automatic stopping of the algorithm, leading to a reduced number of selected noise variables. The new approach is illustrated via simulations and a biomedical real data example.

E1348: **A random forest approach for modeling bounded outcomes**
Presenter: **Moritz Berger**, University of Bonn, Germany
Co-authors: Leonie Weinhold, Matthias Schmid, Richard Mitchell, Kelly Maloney, Marvin Wright

In observational studies one frequently encounters bounded outcome variables, for example, relative frequency measures restricted to the unit interval $(0,1)$. A flexible approach to relate the bounded outcome to a set of explanatory variables is beta regression. In parametric beta regression models one usually assumes that the effects of the explanatory variables on the outcome are linear. In many applications, however, this assumption is too restrictive, for example, when higher-order interactions between the explanatory variables are present. Furthermore, parametric models may not be applicable to high-dimensional data, for example when the number of explanatory variables exceeds the number of observations. To address these issues we propose a random forest approach tailored to the modeling of bounded outcome variables. In contrast to classical random forest algorithms with continuous outcome, which use the mean squared error as splitting criterion, we propose to use the likelihood of the beta distribution for tree building. In each iteration of the tree-building algorithm one chooses the combination of explanatory variable and split point that maximizes the log-likelihood function of the beta distribution, with the parameter estimates directly derived from the nodes of the currently built tree. The method is implemented in the R package ranger.

EO695 Room MAL B04 ROBUST MULTIVARIATE METHODS Chair: Thomas Verdebout

E0475: **Spatial CART classification trees**
Presenter: **Jean-Michel Poggi**, University Paris-Sud Orsay, France
Co-authors: Avner Bar-Hen, Servane Gey

CART (Classification And Regression Trees) is a statistical method designing tree predictors for both regression and classification. We restrict our attention on the classification case with two populations. Each observation is characterized by some input variables gathered in X and a binary response variable Y. The principle of CART is to recursively partition the input space using binary splits and then to determine an optimal partition for prediction. The representation of the model relating Y to X is a tree representing the process of construction of the model. If the explanatory variables are spatial coordinates, we get a spatial decision tree and this induces a tessellation of the input space. We propose a spatial variant of CART method, SpatCART. While usual CART tree considers marginal distribution of the response variable at each node, we propose to take into account the spatial location of the observations. We introduce a dissimilarity index based on Ripley's intertype K-function quantifying the interaction between two populations. This index used for the growing step of the CART strategy, leads to a heterogeneity function consistent with the original CART algorithm. The proposed procedure SpatCART is finally applied to a tropical forest example.

E1429: **An essay on copula modelling for discrete random vectors; or how to pour new wine into old bottles**
Presenter: **Gery Geenens**, University of New South Wales, Australia

Copulas have now become ubiquitous statistical tools for describing, analysing and modelling dependence between random variables. Sklar's theorem, "the fundamental theorem of copulas", makes a clear distinction between the continuous case and the discrete case, though. In particular, the copula of a discrete random vector is not identifiable, which causes serious inconsistencies. In spite of this, downplaying statements are widespread in the related literature, and copula methods are used for modelling dependence between discrete variables. The soundness of copula modelling for discrete data is called to be reconsidered. A more fundamental construction is suggested which allows copula ideas to smoothly carry over to the discrete case. Actually it is an attempt at rejuvenating some century-old ideas of Udny Yule, who mentioned a similar construction a long time before copulas got in fashion.

E1524: **Estimating local rotations**
Presenter: **Agnese Panzera**, University of Florence, Italy
Co-authors: Marco Di Marzio, Stefania Fensore, Charles C Taylor

Nonparametric estimation is considered when the regression function at a point is modeled by a rotation matrix. In particular, we deal with the case where a specific rotation is possible for each location on the surface of a hypersphere. We provide some theoretical properties along with practical applications.

E1872: **Sign tests for weak principal directions**
Presenter: **Julien Remy**, Universita Libre de Bruxelles, Belgium
Co-authors: Davy Paindaveine, Julien Remy, Thomas Verdebout

Inference on the first principal direction of a p-variate elliptical distribution is considered. We do so in challenging double asymptotic scenarios for which this direction eventually fails to be identifiable. In order to achieve robustness not only with respect to such weak identifiability but also with respect to heavy tails, we focus on sign-based statistical procedures, that is, on procedures that involve the observations only through their direction from the center of the distribution. We first focus on weak identifiability setups involving single spikes. We show that, irrespective of the degree of weak identifiability, such setups offer local alternatives for which the corresponding sequence of statistical experiments converges in the Le Cam sense. We exploit this convergence result to build optimal sign tests for the problem considered. In classical asymptotic scenarios where the spectrum is fixed, these tests are shown to be asymptotically equivalent to the sign-based likelihood ratio tests available in the literature. Unlike the latter, however, the proposed sign tests are robust to arbitrarily weak identifiability. We show that our tests meet the asymptotic level constraint irrespective of the structure of the spectrum, hence also in possibly multi-spike setups. We fully characterize the non-null asymptotic distributions of the corresponding test statistics under weak identifiability, which allows us to quantify the corresponding local asymptotic powers.

E1958: **Characterizing multivariate distributions**
Presenter: **Yvik Swan**, Universite de Liege, Belgium
Co-authors: Gesine Reinert, Guillaume Mijoule

A new collection of Stein-type characterizations for multivariate distributions is proposed. We apply these to several well-known families of distributions as well as to some intractable distributions. Applications to Goodness-of-Fit testing and estimation are outlined.

EO184 Room MAL B20 ADVANCES IN CAUSAL INFERENCE METHODS Chair: Yuya Sasaki

E0257: **Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs**
Presenter: **Sebastian Calonico**, Columbia University, United States
Co-authors: Matias Cattaneo, Max Farrell

Modern empirical work in Regression Discontinuity (RD) designs employs local polynomial estimation and inference with a mean square error (MSE) optimal bandwidth choice. This bandwidth yields an MSE-optimal RD treatment effect estimator, but is by construction invalid for inference. Robust bias corrected (RBC) inference methods are valid when using the MSE-optimal bandwidth, but we show they yield suboptimal confidence intervals in terms of coverage error. We establish valid coverage error expansions for RBC confidence interval estimators and use these results to propose new inference-optimal bandwidth choices for forming these intervals. We find that the standard MSE-optimal bandwidth for the RD point estimator must be shrunk when the goal is to construct RBC confidence intervals with the smaller coverage error rate. We further optimize the

constant terms behind the coverage error to derive new optimal choices for the auxiliary bandwidth required for RBC inference. Our expansions also establish that RBC inference yields higher-order refinements (relative to traditional undersmoothing) in the context of RD designs. Our main results cover sharp and sharp kink RD designs under conditional heteroskedasticity, and we discuss extensions to fuzzy and other RD designs, clustered sampling, and pre-intervention covariates adjustments. The theoretical findings are illustrated with a Monte Carlo experiment and an empirical application.

E0259: **Identifying marginal treatment effects in the presence of sample selection**
Presenter: **Otavio Bartalotti**, Iowa State University, United States
Co-authors: Desire Kedagni

Two identification results are developed for the marginal treatment effect (MTE) when there is sample selection. We show that the MTE is partially identified for individuals who are always selected regardless of treatment, and we derive sharp bounds on this parameter under various assumptions. The first identification result combines the standard MTE assumptions with monotonicity of the sample selection variable with respect to the treatment, while the second uses an additional (possibly invalid) instrument. Both results rely on a mixture reformulation of the problem. In the first approach, the mixture weights are identified. We therefore extend a previous trimming procedure to the MTE context. The second identification result relies on the mixture weights varying with the additional instrument, while the mixture component distributions do not.

E0268: **Better bunching, nicer notching**
Presenter: **Marinho Bertanha**, University of Notre Dame, United States
Co-authors: Andrew McCallum, Nathan Seegert

Bunching estimators use mass points in an observed distribution to estimate parameters of a structural model, such as the elasticity of taxable income with respect to tax rates. The distribution of income typically shows mass points at income values with a change in the tax regime: either a change in marginal tax rate (kink) or a change in lump-sum tax (notch). Identification of the elasticity parameter in a setting with both kinks and notches is studied. First, we find that inference methods for the elasticity that focus on one kink may still be valid, as long as there are no notches near that kink. Second, contrary to what was previously thought, it is impossible to identify the elasticity using a kink, when the distribution of agents is non-parametric. We show the same is not true for notches. Third, we propose practical solutions for the lack of identification. We derive partial identification bounds on the elasticity using non-parametric shape restrictions on the distribution of agents. Then, we connect the bunching problem to the literature on censored regressions, namely Tobit and censored quantile regressions. This allows us to combine covariates with semi-parametric restrictions on the distribution of agents to point-identify the elasticity. We compare our estimates to previous estimates based on tax return data in the context of the "earned income tax credit", and find economically meaningful differences.

E0287: **Covariate distribution balance via propensity scores**
Presenter: **Pedro SantAnna**, Vanderbilt University, United States
Co-authors: Xiaojun Song, Qi Xu

The propensity score plays an important role in causal inference with observational data. However, it is well documented that under slight model misspecifications, propensity score estimates based on maximum likelihood can lead to unreliable treatment effect estimators. To address this practical limitation, a new framework is proposed for estimating propensity scores that mimics randomize control trials (RCT) in settings where only observational data is available. More specifically, given that in RCTs the joint distribution of covariates are balanced between treated and not-treated groups, we propose to estimate the propensity score by maximizing the covariate distribution balance. The proposed propensity score estimators, which we call the integrated propensity score (IPS), are data-driven, do not rely on tuning parameters such as bandwidths, admit an asymptotic linear representation, and can be used to estimate many different treatment effect measures in a unified manner. We derive the asymptotic properties of inverse probability weighted estimators for the average, distributional and quantile treatment effects based on the IPS and illustrate their relative performance via Monte Carlo simulations and three empirical applications. An implementation of the proposed methods is provided in the new package `IPS` for `R`.

E0341: **Quantile treatment effects in regression kink designs**
Presenter: **Yuya Sasaki**, Vanderbilt University, United States

The literature on regression kink designs develops identification results for average effects of continuous treatments, average effects of binary treatments, and quantile-wise effects of continuous treatments, but there has been no identification result for quantile-wise effects of binary treatments to date. We fill this void in the literature by providing an identification of quantile treatment effects in regression kink designs with binary treatment variables. For completeness, we also develop large sample theories for statistical inference and a practical guideline on estimation and inference.

EO488 Room MAL B35 RECENT ADVANCES ON JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA Chair: Lang Wu

E0996: **Joint models for longitudinal and survival data: An orthodox best linear unbiased predictor approach**
Presenter: **Renjun Ma**, University of New Brunswick, Canada
Co-authors: Xingde Duan

In medical studies, longitudinal and survival outcomes are frequently collected over time on each of many subjects. As a random effects Cox survival model can be characterized as an auxiliary Poisson random effects model, we can employ our techniques on joint modelling for different types of longitudinal data to handle joint modelling of longitudinal and survival outcomes. An optimal estimation of our model has been developed using orthodox best linear unbiased predictor of random effects. The analysis results do not rely on any distributional assumption of random effects. The approach will be illustrated with real data examples.

E1039: **A variable selection method for the joint modelling of longitudinal and survival data with its application**
Presenter: **Tao Wang**, Yunnan Normal University, China

Although there has been extensive research for joint modelling method of longitudinal and survival data in the last two decades motivated by the requirements of increasingly application and the importance of such joint models has been increasingly recognized, but the research on variable selection method for joint models of longitudinal and survival outcomes with lower computational load is still getting on slowly. We propose a novel Bayesian SCAD variable selection method for semi-parametric joint model which consists of a semi-parametric mixed effects model for longitudinal data and a semi-parametric Cox proportional hazards model for survival data linked through shared random effects. We develop the computational program for such a variable selection method. Simulation studies and real data analysis demonstrate that our method performs well.

E1519: **A joint model for truncated and mixed types of longitudinal and survival data**
Presenter: **Lang Wu**, University of British Columbia, Canada

In the analysis of longitudinal data and survival data, joint models are useful since the longitudinal data and survival data are often strongly associated. In practice, the longitudinal data can be highly complicated, such as being truncated and mixed types of discrete and continuous. We will discuss some recent work to address these data complications in joint models. Another challenge for joint models is computation, since the likelihoods of joint models often involve high-dimensional and intractable integrations. We will also discuss a computationally efficient approximate likelihood method. The models and methods will be applied to the analysis of a recent HIV vaccine dataset.

E1568: **Inferring random change point from longitudinal data subject to left-censoring by segmented mechanistic nonlinear models**
Presenter: **Hongbin Zhang**, CUNY (SPH), United States

Random effects change-point models are commonly used to infer individual-specific time of event that induces trend change of longitudinal data. Linear models are often used before and after the change point. However, linear models may fit the completely observed data well, but may be inappropriate when certain portion of data are censored. In applications such as HIV studies, a mechanistic nonlinear model can be derived for the process based on the underlying data-generation mechanisms and such nonlinear model may provide better "predictions" for the censored values. We propose a random change point model in which we model the longitudinal data by segmented nonlinear mixed effect models and address the left-censoring with the data. We propose a Monte Carol EM based method for the inference. We apply the model on an HIV surveillance data to estimate the time from HIV diagnosis to initiation of antiretroviral therapy (ART) initiation and evaluate the method with simulation to gain insights.

E1579: **Scalable joint modeling of longitudinal and competing risks time-to-event data**
Presenter: **Gang Li**, UCLA, United States

Joint modeling of longitudinal and time-to-event data is useful for longitudinal data analysis with possibly nonignorable missing data and for survival analysis with time-dependent covariates that are intermittently measured and/or with measurement errors. However, current estimation and inference methods for joint models are well known to be computationally complex and costly, which do not scale well even to moderate sample size data. The aim is to improve the computational performance of joint modeling methods by developing novel techniques to exploit some specific structures in fitting a joint model. Numerical simulation results and real data illustrations will be presented.

EO268 Room MAL B36 STATISTICAL LEARNING IN PRACTICE Chair: Alejandro Murua

E0403: **Advances in measuring user learning**
Presenter: **Niall Cardin**, Google UK, United Kingdom

Ads power much of the modern internet. However it is hard to know which ads are the most useful and how many ads to show. Ads blindness, a type of long term user learning, measures changes in user behavior in response to changes in the ads shown. The purpose is to introduce ads blindness, existing techniques to measure ads blindness, followed by discussion of newer experimental methods to better address an array of challenges faced when trying to measure long term user learning.

E0463: **Neural model compression for edge computing**
Presenter: **Vahid Partovi Nia**, Ecole Polytechnique de Montreal / Huawei Noah's Ark Lab, Canada

Deep neural networks is an effective tool for many supervised learning tasks, such as voice recognition, object detection, image classification, etc. In 5G technology many optimization tasks rely on effective user behaviour prediction, in which neural networks play a central role. Neural networks tend to have many parameters, in order of millions, or even billions. Therefore, their deployment on edge devices such as cell phones, smart watches, IoT devices, and wireless base stations is a major challenge. We introduce combination of different strategies to simplify neural networks in order to save memory, computation power, and energy.

E0976: **Predicting geo-localized accidents from usage-based insurance GPS data**
Presenter: **Aurelie Labbe**, HEC Montreal, Canada

Usage Based Insurance (UBI) programs are designed to leverage data-driven technologies by collecting driving data from customers using a Global Positioning System (GPS) and adjust premiums based on driver-level surrogate safety measures (SSMs) of exposure and driving style. The aim is to quantify relationships between SSMs extracted from UBI GPS data and historical crashes at the link-level. From a statistical perspective, crashes typically occur on roadways, which constrains the events to lie along a linear network. In the past years, substantial research efforts have been devoted to the analysis of point processes with the development of methods for point patterns of events that occur on a network of lines. In such models, one can assume that crash coordinates are produced by a Poisson point process whose domain corresponds to edges in the road network. The main challenges posed by the integration of those three sources of data in order to fit point process models on the road network are reviewed: the sparsity of the GPS data and the resulting need to define an imputation strategy for covariates values on the network, limitations imposed by the importation of OSM data and spatial dependence between neighboring points.

E1482: **Car racing strategy**
Presenter: **Stavros Tsalidis**, Quantumblack Mckinsey, United Kingdom

Car racing is a popular competition in which fully or partially electric powered cars race a fixed period of time and/or race laps. The total energy allowed to be used in the race is fixed or restricted and has to be managed wisely in order to win points by overtaking but not run out of energy before end of race. A racing strategy consists in allocating energy budgets dynamically at each lap start for optimal positioning. We build an environment simulating the effects of the energy allocation on the state of a race and use simulations to investigate the outcome of strategies and scenarios in energy budgeting. Parameters of simulations for different drivers are estimated from past races data. We apply reinforcement learning techniques to estimate optimal policies/strategies for energy allocation. The simulation environment is used to validate the estimated optimal strategies.

E1758: **Hyperbolic support vector machines**
Presenter: **Nicolas Wicker**, University of Lille, France
Co-authors: Aya El Dakdouki, Yann Guermeur

Support vector machines aim at separating two classes by a hyperplane. What happens if this surface is changed? We answer this question when hyperplanes are replaced by hyperboloids and show that the obtained class of functions is still learnable and improves in some cases the machine prediction capacities.

EO186 Room MAL G13 TOPICS IN MATHEMATICAL STATISTICS Chair: Natalia A Stepanova

E0392: **Estimating the response density in semiparametric regression**
Presenter: **Ursula Mueller**, Texas A and M University and University of Hamburg, Germany

The focus is on regression models with a parametric (linear or nonlinear) regression function. We assume that the errors have mean zero and are independent of the covariates. The independence assumption enables us to construct a convolution type estimator for the response density that, in general, converges at a faster rate than the usual density estimators. If the regression function is invertible, the estimator converges with the optimal parametric root-n rate. Otherwise the root-n rate cannot be achieved. If the regression function is a step function, we can construct a response density estimator that has the same bias as the usual estimators based on the responses, but a smaller asymptotic variance.

E0540: **Copula-based dynamic models for multivariate time series**
Presenter: **Bouchra R Nasri**, McGill University, Canada
Co-authors: Bruno Remillard

An intuitive way to couple several dynamic time series models even when there are no innovations is proposed. This extends previous work for modeling dependence between innovations of stochastic volatility models. We consider time-dependent and time-independent copula models and we study the asymptotic behavior of some empirical processes constructed from pseudo-observations, as well as the behavior of maximum pseudo-likelihood estimators of the associated copula parameters. The results show that even if the univariate dynamic models depend on unknown

parameters, the limiting behavior of many processes of interest does not depend on the estimation errors. One can perform tests for change points on the full distribution, the margins or the copula as if the parameters of the dynamic models were known. This is also true for some parametric models of time-dependent copulas. This interesting property makes it possible to construct consistent tests of specification for the dependence models, without having to consider the dynamic time series models. Monte Carlo simulations are used to demonstrate the power of the proposed goodness-of-fit test in finite samples. An application to Moroccan hydro-climatic data is given.

E0705: **On estimation of the amount of sparsity in normal mixture models**
Presenter: **Natalia A Stepanova**, Carleton University, Canada
Co-authors: Yibo Wang

The motivation comes from a variable selection problem in sparse normal mixtures. For this problem, the sharp selection boundaries, that is, the necessary and sufficient conditions for the possibility of successful variable selection in the exact and almost full regimes are available. The existing selection boundaries, as well as the procedure that provides almost full selection, depend on the fraction of nonzero means, which is generally unknown. We present a new estimator for the fraction of nonzero means in normal mixture models with relatively few nonzero means that are only moderately large. We show that, in the region where variable selection is possible, the new estimator dominates (in terms of the minimax rate of convergence) the existing estimators proposed earlier in similar contexts; the same conclusion continues to hold for the region where signal detection is possible. Moreover, our estimator nearly attains the optimal rate of convergence. The obtained analytical results are illustrated numerically.

E0779: **Optimal tests for unordered paired observations**
Presenter: **Laura Dumitrescu**, Victoria University of Wellington, New Zealand
Co-authors: Laura Dumitrescu

Within a nonparametric framework, the problem of testing the equality of marginal distributions for samples of bivariate data, with unobservable order in each pair, is considered. The approach is based on a symmetrized empirical process and it is shown what the loss is (due to the restriction of observability) in terms of the power of the tests. Furthermore, linear statistics that are asymptotically optimal for testing the equality of marginal distributions against contiguous alternatives are obtained. The advantage of the proposed approach is that it leads to whole a class of test statistics which are asymptotically distribution free, but also, that local alternatives of dependence can be detected.

E0899: **Estimations based on resampling by stable motion**
Presenter: **Manfred Denker**, Penn State University, PA, USA, United States

A nonparametric parameter estimation with large or infinite variance is proposed which is based on stochastic integration and resampling by stable motions. A simulation study and an application to estimating the coupling strength in neural networks is provided.

EO052 Room MAL G15 BAYESIAN MODELS FOR COMPLEX DEPENDENCE STRUCTURES Chair: Alessandra Guglielmi

E0828: **Random partition distribution concentrated around a focal partition**
Presenter: **David Dahl**, Brigham Young University, United States
Co-authors: Richard Warr, Thomas Jensen

Random partition models, such as the Chinese restaurant process, allow a Bayesian model to flexibly borrow strength. While many partition priors are exchangeable, we propose a nonexchangeable prior based on a focal partition, a Bayesian's prior guess for the unknown partition. We show how our approach modifies the Chinese restaurant process so that partitions that are similar to the focal partition have higher probability. There is a weight parameter that varies between 0 and infinity, where 0 corresponds to the original Chinese restaurant process and infinity yields a point mass distribution at the focal partition. In motivation and spirit, our approach is similar to another recent one. In contrast, however, we have a tractable normalizing constant so inference can easily be made on the weight and mass parameter. We investigate the similarity and difference between these approaches.

E0930: **CAR neutral to the right processes**
Presenter: **Federico Bassetti**, Politecnico Milano, Italy
Co-authors: Roberto Casarin, Ilenia Epifani

A class of time dependent Neutral to the Right Processes, CAR-NTR, is introduced. These processes are obtained starting from a Markov sequence of Completely Random Measures, CAR-CRM. By using this processes we obtain a Bayesian non parametric hidden Markov model for positive random variables. We prove various theoretical properties of CAR-CRM and CAR-NTR, in particular we derive closed form expressions for the filtered distribution of the underling process given the past observations.

E1190: **Modelling ethnic differences in metabolic associations via dynamic Bayesian nodewise regression**
Presenter: **Maria De Iorio**, UCL, United Kingdom

A novel approach is proposed to the estimation of multiple Gaussian Graphical Models to analyse dynamic evolving patterns of association among a set of metabolites over different groups of patients. The motivating application is the Southall And Brent REvisited study, a tri-ethnic cohort study conducted in the UK. We are interested in identifying potential ethnic differences in metabolite levels and associations, with the aim of gaining a better understanding of different risk of cardio-metabolic disorders across ethnicities. We model the inverse-covariance structure of a set of metabolites measured over different time points and for three ethnic groups. We adopt a Bayesian adaptation of the Nodewise Regression technique to infer the structure of the graphs. We assume a global-local shrinkage prior over the regression parameters to impose a sparse structure on the graph, and we extend the prior to allow borrowing of information across different groups. Finally, we extend the model to a dynamic framework to impose a time dependence and estimate multiple graphs over time. Posterior inference is performed through Markov Chain MonteCarlo methods. Specifically, we use the software Stan, which employs Hamiltonian Monte Carlo. The proposed approach is able to capture a wide range of graph topologies and identify diverse/common structures across multiple graphs, corresponding to different ethnicities and allows us to estimate time trends for each metabolite connection.

E1229: **Spatio-temporal random partition models**
Presenter: **Garritt Page**, Brigham Young University, United States
Co-authors: Fernando Quintana, David Dahl

The number of scientific fields that regularly collect data that are temporally and spatially referenced continues to rapidly growth. An intuitive feature of spatio-temporal data is that measurements taken on experimental units near each other in time and space tend to be similar. As such, many methods developed to accommodate spatio-temporal dependent structures attempt to borrow strength among units close in space and time, which constitutes an implicit space-time grouping. Rather than implicitly performing this spatio-temporal grouping, we develop a class of dependent random partition models that explicitly models spatio-temporal clustering. Our model is a joint distribution for a sequence of random partitions indexed by time and space. We first detail how temporal dependence is incorporated so that partitions evolve gently over time. Then conditional and marginal properties of the joint model are derived. We then demonstrate how space can be integrated. Computation strategies are detailed and we illustrate the methodology through simulations and applications.

E1728: **Bayesian covariance structure modeling of high-dimensional dependence structures**
Presenter: **Jean-Paul Fox**, University of Twente, Netherlands

The covariance structure of response data represents interesting phenomena such as the inter and intra-individual variability, and response dependencies due to higher-level clusters. The well-known psychometric modeling frameworks (e.g., structural equation modeling, item response theory, multilevel modeling), use latent variables (random effects) to model the covariance structures. However, latent variables have several disadvantages. They demand larger sample sizes, increase the number of model parameters, and limit the flexibility of the model to describe high-dimensional data. To avoid the problems associated with latent variables, the covariance structure is modeled directly by defining conjugate priors for the (co)variance parameters, where a multivariate distribution is defined for the response data. The priors include restrictions on the parameter space of the covariance parameters such that any combination of covariance parameters leads to a positive definite covariance matrix. The priors give support to testing the presence of random effects, reduce boundary effects by allowing non-positive (co)variance parameters, and support accurate estimation even for very small true variance parameters. The priors lead to efficient posterior computation using Gibbs sampling. The advantages of Bayesian Covariance Structure Modeling (BCSM) are illustrated through the joint modeling of high-dimensional response data and process data.

EO522 Room MAL G16 STATISTICAL INFERENCE OF COPULA MODELS Chair: Jean-David Fermanian

E0700: **Multiple change-point for semiparametric copula models**
Presenter: **Olivier Lopez**, Sorbonne Universite Paris, France

A dynamic copula model is considered. We assume that the dependence structure between some random variables is a copula belonging to a fixed parametric copula family, but with association parameter $\theta(t)$ evolving with time t. A multiple change-point model consists in assuming that the function $\theta(t)$ is piecewise constant, without pre-determining the times where the jumps are located. We derive finite sample bounds for maximum likelihood estimation of these times and amplitudes of jumps, and show the consistency of model selection procedures to select the appropriate number of changes.

E0541: **Semiparametric inference for copulas of mixed data**
Presenter: **Bruno N Remillard**, HEC Montreal, Canada
Co-authors: Bouchra R Nasri, Christian Genest, Johanna Neslehova

Inference methods are proposed for the estimation of the parameter of a copula family when the unknown marginal distributions are mixtures of discrete and absolutely continuous distribution functions. Under smoothness assumptions, the estimation errors are shown to be Gaussian and their variance can be estimated.

E0605: **Testing constant conditional dependence structure over partition sets**
Presenter: **Aleksey Min**, Technische Universitaet Muenchen, Germany
Co-authors: Jean-David Fermanian

Recently, the simplifying assumption for conditional copulas has been rigorously investigated and classified, and several testing procedures have been proposed. We consider one particular test of a constant conditional dependence structure over a given partition sets and propose a simple alternative testing procedure based on the equality of Kendall's rank correlation coefficients. The performance of the proposed test will be illustrated in a simulation study.

E0664: **On the estimation of elliptical copula generators**
Presenter: **Alexis Derumigny**, University of Twente, Netherlands
Co-authors: Jean-David Fermanian

The problem of estimating elliptical copula models is studied. They are defined by a correlation matrix and an unknown function (the "generator"). Although the estimation of the correlation matrix has been well-studied, estimating the generator is much harder. This model is in fact not identifiable, in the sense that different generators can correspond to the same copula. Therefore, we propose a possible identification constraint. We present an iterative algorithm to nonparametrically estimate the generator, using kernel-smoothing. Finally, we give some numerical results on simulated and real data.

E0193: **A classification point-of-view about conditional Kendall tau**
Presenter: **Jean-David Fermanian**, Ensae-Crest, France
Co-authors: Alexis Derumigny

The purpose is to show how the problem of estimating conditional Kendall's tau can be rewritten as a classification task. The conditional Kendall's tau is a conditional dependence parameter which can be interpreted as a characteristic of a given pair of observations. The goal is to predict whether the pair is concordant or discordant conditionally on some covariates. We prove consistency and asymptotic normality of a family of penalized approximate maximum likelihood estimators, including the equivalent of the logit and probit regressions in our framework. Then, we detail specific algorithms adapting usual machine learning techniques, including nearest neighbors, decision trees, random forests and neural networks, to the setting of the estimation of conditional Kendall's tau. A small simulation study compares their finite sample properties. Finally, we apply all these estimators to a dataset of European stock indexes.

EO250 Room Chancellor's Hall RECENT DEVELOPMENTS IN TIME SERIES FORECASTING Chair: James Taylor

E0844: **Probabilistic forecasting of an air quality index**
Presenter: **James Taylor**, University of Oxford, United Kingdom
Co-authors: Jooyoung Jeon, Xiaochun Meng

Air pollution has emerged as a major issue affecting human health, with the resulting burden on health systems having economic and political implications. Urban air pollution is believed to be the cause of more than a million premature deaths worldwide each year. Respiratory illnesses are the main health risk factor, but the prevalence of heart disease, stroke and cancer is also increased. Air quality indices are widely-used to summarise the severity of the level of a set of pollutants, with a traffic light signal often used to provide a visual indicator. The index is a convenient measure used by policy makers, but is also used, on a day-by-day basis, by health professionals and the public, especially those with a history of respiratory conditions. Forecasts of the index are typically produced each day for lead times up to several days ahead. The predictions are usually provided by meteorologists using atmospheric models that have chemistry features incorporated. Probabilistic forecasting from such models is not straightforward, and hence is very rare. We consider the use of time series models to produce density forecasts for the index. The approach involves the fitting of a multivariate model to a set of six pollutants. To capture the dependencies between the pollutants in a practical way, we use an empirical copula. The empirical work uses hourly data from South Korea, where more than half the population are considered to be exposed to dangerous levels of pollutants.

E0825: **Regularized regression for hierarchical forecasting without unbiasedness conditions**
Presenter: **Souhaib Ben Taieb**, University of Mons, Belgium
Co-authors: Bonsoo Koo

Forecasting a large set of time series with hierarchical aggregation constraints is a central problem for many organizations. However, it is particularly challenging to forecast these hierarchical structures. In fact, it requires not only good forecast accuracy at each level of the hierarchy, but also the coherency between different levels, i.e., the forecasts should satisfy the hierarchical aggregation constraints. Given some incoherent base

forecasts, the state-of-the-art methods compute revised forecasts based on forecast combination which ensures that the aggregation constraints are satisfied. However, these methods assume the base forecasts are unbiased and constrain the revised forecasts to be also unbiased. We propose a new forecasting method which relaxes these unbiasedness conditions, and seeks the revised forecasts with the best tradeoff between bias and forecast variance. We also present a regularization method which allows us to deal with high-dimensional hierarchies, and provide its theoretical justification. Finally, we compare the proposed method with the state-of-the-art methods both theoretically and empirically. The results on both simulated and real-world data indicate that our methods provide competitive results compared to the state-of-the-art methods.

E0905: **Probabilistic forecasting of patient waiting times in the emergency department**
Presenter: **Siddharth Arora**, University of Oxford, United Kingdom
Co-authors: James Taylor

Accurate estimates of patient waiting times in the emergency department (ED) have been associated with increased patient satisfaction and improved outcomes. Moreover, waiting time estimates can assist hospitals to streamline patient-flow based on informed staff and resource allocation. Individual patient waiting times in ED are inherently uncertain. We thus generate and evaluate probabilistic forecasts, based on the following categories of predictor variables: (1) workload, (2) staffing, (3) calendar variables, (4) demographics, and (5) severity of the patient condition. Using around 350,000 anonymized patient-level ED records collected over a period of five years for one of the major hospital sites in the UK, we develop a methodology to: (1) predict patient waiting times for both major and minor triage categories, (2) identify the variables that have the highest impact on modelling accuracy, and (3) accommodate the dynamic nature of patient-flow in ED via re-estimation of predictor variables. Out-of-sample point and probability distribution forecasts are evaluated using the mean absolute error (MAE) and continuous ranked probability score (CRPS), respectively.

E1623: **A Bayesian long short-term memory model for value at risk and expected shortfall joint forecasting**
Presenter: **Zhengkun Li**, The University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, Richard Gerlach, Junbin Gao

Value at Risk (VaR) and Expected Shortfall (ES) are widely used by financial institutions to measure the market risk and avoid the extreme market movement. The proposal of Asymmetric Laplace (AL) score function allows to backtest VaR and ES forecasting accuracy jointly, as well as the joint modelling of VaR and ES by constructing the optimization problem with the AL score function entered as the underlying objective function. As the result, several linear models and their extensions were proposed. However, it is still challenging to model the possible nonlinear factors in the underlying dynamics. We address the problem by developing a Bayesian approach that can capture the nonlinear and long-term effects with the Long Short-Term Memory (LSTM) structure, a neural network approach for time series modelling. We further adapt the adaptive Markov chain Monte Carlo (MCMC) algorithm to estimate the model parameters based on the AL log likelihood function. Empirical results show that the proposed Bayesian LSTM-AL model can improve the forecasting accuracy over a range of well-established and recent proposed models, the advantages can be observable from both test results and dynamic plots.

E0858: **Scoring functions for forecasts of multivariate distributions and level sets**
Presenter: **Xiaochun Meng**, University of Sussex, United Kingdom
Co-authors: James Taylor, Souhaib Ben Taieb, Siran Li

Interest in the prediction of multivariate probability distributions is growing due to the increasing availability of rich datasets and computational developments. Scoring functions enable the comparison of forecast accuracy, and can potentially be used for estimation. A scoring function for multivariate distributions that has gained some popularity is the energy score. This is a generalization of the continuous ranked probability score (CRPS), which is widely used for univariate distributions. A little-known, alternative generalization is the multivariate CRPS (MCRPS). We propose a new theoretical framework for scoring functions for multivariate distributions, which encompasses the energy score and multivariate CRPS as specific cases. This framework can be used to generate new scores, and we demonstrate this with the introduction of a score based on the density function. For univariate distributions, it is well-established that the CRPS can be expressed as the integral over the quantile check loss score. We show that, in a similar way, scoring functions for multivariate distributions can be disintegrated to obtain the scoring functions for three types of level sets: projection quantiles, isoprobabilty contours and density contours. To compute the various scoring functions, we propose a simple numerical algorithm. We use a simulation study to support our findings.

EO624 Room CLO 101 STATISTICAL METHODS FOR BEHAVORIAL DATA Chair: Philip Reiss

E0314: **Single paper meta-analysis**
Presenter: **Blake McShane**, Northwestern University, United States
Co-authors: Ulf Bockenholt

A typical behavioral research paper features multiple studies of a common phenomenon that are analyzed solely in isolation. Because the studies are of a common phenomenon, this practice is inefficient and forgoes important benefits that can be obtained only by analyzing them jointly in a single-paper meta-analysis (SPM). To facilitate SPM, we introduce meta-analytic methodology that is user-friendly, widely applicable, and specially tailored to the SPM of the set of studies that appear in a typical behavioral research paper. Our SPM methodology provides important benefits for study summary, theory testing, and replicability that we illustrate via several case studies. We advocate that authors of typical behavioral research papers use it to supplement the single-study analyses that independently examine the multiple studies in the body of their papers as well as the qualitative meta-analysis that verbally synthesizes the studies in the general discussion of their papers. When used as such, this requires only a minor modification of current practice. We provide an easy-to-use website that implements our SPM methodology.

E0408: **The actor-partner interdependence model for longitudinal dyadic data in the SEM-framework**
Presenter: **Tom Loeys**, Ghent University, Belgium

Many of the phenomena studied by behavioral scientists are interpersonal by definition. While historically behavioral data were mostly gathered on individuals, these are unprecedented times in terms of the availability of high-quality dyadic. When two people interact in a relationship, the outcome of each person can be affected by both his or her own inputs and his or her partners inputs. The Actor-Partner Interdependence Model (APIM) offers an appealing approach to model such data. When one collects repeated measures on dyads, one must not only contend with the non-independence of the members within a dyad, but also the correlation of the longitudinal measures within a dyad member. This can be achieved by either modeling an autoregressive residual covariance structure or by including a lagged dependent variable in the mean structure. The implementation of the first approach is readily available in multilevel software, but is lacking in the SEM-framework. A complication for the second approach lies in noting that the combination of a lagged outcome in the mean structure and a random intercept can lead to biased inference within standard multilevel software due to the exogeneity assumption. Solutions to those statistical and computational challenges are presented.

E0712: **Generalized reliability based on distances**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Meng Xu, Ivor Cribben

The intraclass correlation coefficient (ICC) is a classical index of measurement reliability. With the increasing prevalence of complex data objects such as curves, images or graphs for which the ICC is not defined, there is a need for new ways to assess reliability. To meet this need, a generalization of the ICC, defined in terms of arbitrary distances among observations, is proposed. The Spearman-Brown formula, which shows how more intensive measurement increases reliability, is extended to encompass the distance-based ICC. A simple bias correction is proposed to

improve the coverage of bootstrap confidence intervals for the (classical or distance-based) ICC, and its efficacy is demonstrated via simulations. The proposed methodology is illustrated by analyzing the test-retest reliability of brain connectivity networks derived from a functional magnetic resonance imaging study.

E0977: **Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model**
Presenter: **Laura Bringmann**, University of Groningen, Netherlands
Co-authors: Casper Albers

Recent studies have shown that emotion dynamics such as inertia (i.e., autocorrelation) can change over time. Importantly, current methods can only detect either gradual or abrupt changes in inertia. This means that researchers have to choose a priori whether they expect the change in inertia to be gradual or abrupt. This will leave researchers in the dark regarding when and how the change in inertia occurred. Therefore, a new model is used: the time-varying change point autoregressive (TVCP-AR) model. The TVCP-AR model can detect both gradual and abrupt changes in emotion dynamics. More specifically, this shows that the inertia of positive affect and negative affect measured in one individual differ qualitatively in how they change over time. Whereas the inertia of positive affect increased only gradually over time, negative affect changed both in a gradual and abrupt fashion over time. This illustrates the necessity of being able to model both gradual and abrupt changes in order to detect meaningful quantitative and qualitative differences in temporal emotion dynamics.

E1091: **Recursive partitioning of clustered and longitudinal data with GLMM trees**
Presenter: **Marjolein Fokkema**, Leiden University, Netherlands

Subgroup and moderator detection is of interest in many research fields, e.g., to find subgroups that show differences in treatment effects, or differences in growth trajectories over time. Recursive partitioning or decision-tree methods are pre-eminently suited for this task. Often, researchers may want to detect subgroups or moderators in clustered or longitudinal data. Several existing recursive partitioning methods allow for detecting subgroups in clustered and longitudinal data, like SEM trees, RE-EM trees and GLMM trees. These methods estimate a recursive partition, while accounting for dependence between observations through the estimation of random effects. At the same time, the methods differ in terms of model specification and estimation procedures. Differences and similarities between the methods are discussed. Furthermore, the focus will be on the different ways in which GLMM trees can be specified, and how characteristics of the data problem can best be accounted for, like the measurement level of the partitioning variables, or the strength of the random effects. Simulation results will be presented, and a real data example on subgroup detection in children reading trajectories will be used to illustrate the effects of different model specifications and estimation procedures.

EO074 Room Court CLUSTERING OF MULTIVARIATE DEPENDENT DATA Chair: Marta Nai Ruscone

E0814: **Predicting disease progression in neurodegenerative diseases with high phenotypic variability**
Presenter: **Frank Dondelinger**, Lancaster University, United Kingdom

Identifying factors that influence the clinical progression of neurodegenerative diseases is of critical importance to both experimentalists trying to understand the disease mechanisms, and clinical researchers trying to develop improved therapies. While much effort has gone into the detection of risk factors for a given disease, most of these approaches ignore the inherent variability in the clinical phenotypes. We have developed a high-dimensional mixture model approach for jointly solving the problem of data-driven estimation of clinical phenotypes and prediction of disease progression. Longitudinal dynamics are captured via a mixed model approach, and we take into account both the distribution of the response and the distribution of the covariates for estimating the disease phenotypes. We demonstrate the performance of our method by applying it to data from the PROACT database on amytrophic lateral sclerosis, as well as data from the Alzheimers Disease Neuroimaging Initiative (ADNI). We show that in both cases joint inference of the subtypes and predictors improves the prediction performance, and hence the clinical usefulness of our results.

E0653: **Semiparametric copula-based mixture models**
Presenter: **Gildas Mazo**, INRA, France

Faced with non-Gaussian clusters (or more generally, non-elliptical clusters), natural alternatives to Gaussian mixture models include copula-based mixture models and nonparametric mixture models. The first can be difficult to calibrate. The last are built on a conditional independence assumption. With semiparametric copula-based mixture models, one aims at getting benefits from both approaches: exploiting the underlying dependence structure and getting some nonparametric flexibility. Semiparametric copula-based mixture models are presented and algorithms are given to do the inference. These are illustrated on simulated and real data.

E0180: **Nonparametric clustering for spatio-temporal data**
Presenter: **Ashwini Venkatasubramaniam**, Alan Turing Institute, United Kingdom
Co-authors: Konstantinos Ampountolas, Ludger Evers

A non-parametric clustering approach for spatio-temporal dataset is proposed which seeks to identify spatially contiguous clusters and retain the underlying temporal patterns for associated clusters. This flexible Bayesian method utilises a modified distance dependent Chinese restaurant process (ddCRP), referred to as the netCRP, to model network connectivity and spatial dependencies. The netCRP seeks to incorporate neighbourhood relationships for each vertex in the graph and the graph network in this context is assumed to be composed of vertices that have a limited number of adjacent vertices. The non-sequential ddCRP is modified to also allow for the ability to control the number of self-links and redundant links between vertices; individual clusters in the network are formed by the formation of cycles. In order to fully account for within-cluster spatial and temporal correlation, the model defines a spatio-temporal precision matrix using a type of conditional auto-regressive (CAR) model and first order auto-regressive (AR-1) model. The model utilises a Metropolis within Gibbs sampler to fully explore all possible cluster configurations in the network and infer the relevant parameters. We illustrate this developed clustering method using applications to grid-style and map-based graph networks.

E1304: **Comparing EM to a greedy search algorithm to optimize ICL for mixture models**
Presenter: **Arthur White**, Trinity College Dublin, Ireland
Co-authors: Gilles Celeux, Jason Wyse

The integrated complete-data likelihood (ICL) is a popular criterion in model-based clustering for choosing the number of clusters of a finite mixture model. Typically, the ICL is computed using a BIC-like approximation, which depends on maximum likelihood estimates that are found using the expectation-maximisation (EM) algorithm. An alternative method for clustering with the ICL calculates the exact ICL in closed form within a Bayesian framework. A greedy search (GS) algorithm is then used to allocate observations to clusters in order to maximise the ICL directly and hence obtain an optimal clustering solution. This approach can be used to simultaneously search the model space and cluster the data. To better understand the properties of the GS method, we conducted an extensive simulation study comparing its performance to the standard EM approach, in terms of number of clusters selected, cluster accuracy, and computational cost. The performance of the methods on real data is also discussed.

E1833: **Model-based clustering of longitudinal data with nominal, multidimensional outcomes**
Presenter: **Marc Scott**, New York University, United States
Co-authors: Kaushik Mohan, Jacques-Antoine Gauthier

Methods and models for longitudinal data with multidimensional nominal outcomes are somewhat limited, but they are essential to the study of the life course. In that domain, interest centres on the time and order of life course events such as having children and working full or part-time and the duration of the phases that they delineate. Typical behaviour, or typologies, are often desired, and clustering using optimal string matching

algorithms or parametric models of duration in a competing risks framework are used; the appropriateness of each derives from competing goals and orientation. We focus on the latter, model-based approach to clustering dependent data such as this, positing a parsimonious data generating process (DGP) with a novel error structure. This provides us with the ability to: simulate individual trajectories; modify trajectory characteristics over time by conditioning on variables; handle multi-state trajectories and missing outcomes. Several of these goals are particularly challenging when the number of states is of moderate size and many transitions are infrequent and/or time in-homogeneous. Using the Swiss Household Panel (SHP), we demonstrate the appropriateness of a class of clustering models for sequences with heterogeneous dependence structure that provide new techniques for assessing goodness of fit as well as yield insights into social processes.

EO314 Room MAL 152 MODEL SPECIFICATION TESTS Chair: Maria Dolores Jimenez-Gamero

E0598: **Testing the equality of a large number of populations**
Presenter: **Marta Cousido Rocha**, University of Vigo, Spain
Co-authors: Maria Dolores Jimenez-Gamero, Virtudes Alba-Fernandez

Given k independent samples with finite but arbitrary dimension, the problem of testing for the equality of their distributions, that can be continuous, discrete or mixed, is considered. In contrast to the classical setting where k is assumed to be fixed, and the sample size from each population increases without bound, k is assumed to be large, and the size of each sample is either bounded or small in comparison to k. The asymptotic distribution of the considered test statistic is stated under the null hypothesis of equality of the k distributions, as well as under alternatives, which let us study the consistency of the resulting test. Specifically, it is shown that the proposed test statistic is asymptotically free distributed under the null hypothesis. The finite sample performance of the test based on the asymptotic null distribution is studied via simulation.

E0630: **Goodness-of-fit tests for the geometric distribution based on conditional moments**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Maria Dolores Jimenez-Gamero

The geometric distribution is a discrete law which is popular in life time analysis as a discrete survival distribution. Recently, nice applications in capture-recapture methods have been found. A test is proposed for testing goodness-of-fit to that distribution. The test is based on a characterization of the geometric distribution in terms of the conditional expectation of the second order statistic, given the value of the first order statistic, which is linear if and only if the law is geometric. Then, using the well-known Bierens characterization of conditional moments, a test statistic is proposed. The asymptotic null distribution of the test statistic under general conditions is derived. A parametric bootstrap can be used to consistently approximate the null distribution. The goodness of the bootstrap estimator for finite sample sizes is assessed by simulation. The power of the new test is numerically compared with that of other existing tests, concluding that the proposal shows a competitive behavior.

E1041: **Test for detecting risk equivalent or risk neutral portfolios**
Presenter: **Daniel Gaigall**, Leibniz University Hannover, Germany
Co-authors: Marc Ditzhaus

We are interested in testing the marginal homogeneity or the exchangeability of the joint distribution of two portfolios. Thereby, we deal with the underlying time series of the portfolios as functional data. The test under consideration is of Cramér-von-Mises type and is applicable for random variables in a general Hilbert space. We suggest a bootstrap procedure to obtain critical values. From the theory of U-statistics, properties of the test under the null hypothesis and under alternatives are derived. A direct application is the detection of risk equivalent or risk neutral portfolios. We analyze real financial time series to demonstrate how the approach works in practice.

E1244: **New tests of multivariate normality by a characterizing property of the Hermite operator**
Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany
Co-authors: Norbert Henze

A novel class of affine invariant tests for normality in any dimension is proposed. The starting point is a characterization of multivariate normality by a partial differential equation motivated by the Hermite operator and its eigenfunctions. The test statistic thus uses the second partial derivative of the empirical characteristic function. Weak convergence results under the null, under fixed and under contiguous alternatives are derived. A finite sample Monte Carlo simulation study shows that the new statistic outperforms most of the well established procedures.

E1414: **Testing for superiority between two variance functions**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Graciela Boente

The focus is on the problem of testing the null hypothesis that the variance functions of two populations are equal versus one-sided alternatives under a general nonparametric heteroscedastic regression model. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root-n contiguous alternatives. A Monte Carlo study is performed to analyse the finite sample behaviour of the proposed test.

EO580 Room MAL 153 STATISTICAL METHODS FOR PRECISION MEDICINE Chair: Paul Kirk

E0764: **Relating clusters of patients to distal outcomes: A misclassification-correction approach**
Presenter: **Yajing Zhu**, MRC Biostatistics Unit, University of Cambridge, United Kingdom
Co-authors: Steven Kiddle

Latent class analysis (LCA) is widely used to derive categorical variables from multivariate data. One particular interest is to relate these classes to distal outcomes. Commonly used approaches include (1) the one-step approach, where the mixture model and the regression model are estimated simultaneously and (2) naive step-wise approach (e.g. the modal class approach) where subjects are first assigned to the latent class with the highest posterior probability and then treated as observed predictor in the regression model. However, regression coefficients for the modal class will be biased due to the unintended influence of the distal outcome on class membership (the one-step approach) and the misclassification error (the modal class approach). To address these problems, we treat the derived classes as an imperfect measurement of the true class in the regression for the distal outcome, with measurement error determined by the misclassification probabilities. This misclassification-correction approach is applied to a study using the UK electronic health records to understand the impact of multimorbidity (co-existence of more than one long term conditions) on service use and mortality.

E1352: **On high-dimensional prediction for diagnostics**
Presenter: **Bernd Taschler**, German Center for Neurodegenerative Diseases, Germany
Co-authors: Sach Mukherjee

The purpose is to discuss questions that arise in the use of high-dimensional prediction methods in diagnostic applications, illustrated by a case study in leukaemia. We focus, in particular, on machine learning and sparse regression approaches that learn high-dimensional predictive signatures directly from genome-wide data. We discuss issues arising from batch and site effects, the importance of understanding the effect of prevalence in the tested population and on the need for careful empirical assessment of predictors.

E0878: **Scaling up nonparametric Bayesian clustering with MCMC for big data applications**
Presenter: **Boris Hejblum**, Universita de Bordeaux, Inserm BPH U1219, Inria SISTM, Vaccine Research Institute, France
Co-authors: Paul Kirk

Non-parametric Bayesian mixture models such as Dirichlet process mixture models (DPMMs), can be used to perform model-based clustering. One of their advantages is their ability to directly estimate the number of clusters from the data, avoiding the tricky issue of choosing the number of clusters. While state-of-the-art Markov chain Monte Carlo (MCMC) algorithms allow efficient and exact inference for these DPMM, it is generally difficult to scale up these algorithms to hundreds of thousands of data points. Subsampling approaches present several drawbacks, especially when some clusters are quite rare. We propose instead a two-step strategy: (i) first summarize the dataset using a misspecified, largely over-parametrized but simple clustering algorithm (such as k-means); and then (ii) use the resulting weighted summarization of the dataset to perform Bayesian inference for the DPMM via MCMC algorithms. We use numerical simulations as well as real single-cell cytometry data to investigate the properties of this strategy.

E1307: **A Bayesian clustering approach for the analysis of repeated cognitive tests, imaging, genetic data and time-to-dementia**
Presenter: **Anais Rouanet**, Bordeaux Population Health Center, France
Co-authors: Sylvia Richardson, Paul Kirk, Brian Tom

The aim is to present an outcome-guided Bayesian Dirichlet Process Mixture Model developed to identify patterns of cognitive decline associated with differential dementia risk, as well as specific profiles of baseline socio-demographic, imaging and genetic data. Our model links a longitudinal outcome, a time-to-event and a set of correlated variables, called profile variables, to help identify clusters of patients. Adopting a Bayesian approach with a Dirichlet Process prior upon the mixture distribution allows us to quantify the uncertainty in both the number of clusters and their characteristic profiles. Given the cluster allocations, the longitudinal outcome and the profile variables are described by mixed-effects model and multinomial (or multivariate Gaussian) distributions, respectively. The model is estimated by MCMC and uncertainty of the final partition is assessed through model-averaging techniques. We present results from the North American Alzheimer's Disease NeuroImaging (ADNI) study, which demonstrate the utility of our approach to refine the stratification of subjects with high risk of dementia.

E1235: **Individual-level prediction of Alzheimers disease progression: Tackling the TADPOLE challenge**
Presenter: **Steven Hill**, MRC Biostatistics Unit, University of Cambridge, United Kingdom
Co-authors: James Howlett, Steven Kiddle, Sach Mukherjee, Anais Rouanet, Bernd Taschler, Brian Tom, Simon White

Alzheimers disease is an increasing burden on public healthcare systems. It is thought that treatments are most likely to be effective in the very early stages of the disease process. However, identifying individuals in these early stages is challenging. The Alzheimers Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge aimed to investigate the ability of methods to predict the future progression of individuals at risk of Alzheimers disease. Training and test data for this open community challenge was from the Alzheimers Disease Neuroimaging Initiative (ADNI) and the task was to predict three outcomes: clinical diagnosis, a cognitive test score and a brain imaging biomarker. We used several approaches to tackle this problem, including multi-state models, latent-class mixed models and high-dimensional regression. Overall, our methods performed well and were top-ranking in the cross-sectional prediction subcategory. An overview of the challenge, our approaches and their performance will be provided, and some insights from our participation in the challenge will be drawn.

EO669 Room Senate EXTREMES, DEPENDENCIES AND APPLICATIONS Chair: Katharina Hees

E1170: **Patterns in spatio-temporal extremes**
Presenter: **Marco Oesting**, University of Siegen, Germany
Co-authors: Alexander Schnurr

In many applications in environmental sciences, extreme events exhibit a complex spatial-temporal structure which needs to be described in a compressed way. To this end, we propose to investigate patterns originating from the temporal course of various characteristics of the extreme event. Examples include the magnitude of the event, its centroid or the size of the affected area. We verify the existence of the corresponding limit distributions in the framework of regular variation, develop non-parametric estimators and show their asymptotic normality under appropriate mixing conditions. The finite-sample behaviour will be demonstrated in simulated and real data examples.

E1405: **Estimating return levels from serially dependent observations**
Presenter: **David Walshaw**, Newcastle University, United Kingdom
Co-authors: Lee Fawcett

A practical overview of the problem of estimation of return levels for serially dependent observations is taken. Starting with the basic idea of an Exceedances-Over-Thresholds analysis, we consider the range of options available for dealing with the clustering of exceedances brought about by the temporal dependence in the process. We discuss the bias incurred by using a Peaks-Over-Thresholds (POT) approach, and the various alternatives for addressing this. Simply using all exceedances removes the bias but underestimates the variance of estimators. This can be dealt with by using composite likelihood based methods to inflate the variance estimates. However more sophisticated methods based on estimating the extremal index which characterizes the strength of extremal clustering enable us to produce model-based estimators which perform well, and can also be employed in the context of Bayesian inference.

E0350: **A time series model for global horizontal irradiation based on extreme value theory and copulas**
Presenter: **Alfred Mueller**, University of Siegen, Germany
Co-authors: Matthias Reuber

The importance of renewable energies, especially photovoltaics, in the worldwide electricity generation has increased over the past years. Thus, there is an increasing demand for probabilistic hourly models for local and global PV yields. We use an indirect modeling approach of local PV yields with irradiation data provided by the Copernicus Atmosphere Monitoring Service. We propose a statistical estimation for lower and upper bounds of global horizontal irradiations based on extreme value theory. Moreover, we introduce copula based time series models for the hourly and daily dependence structure. Time dependent parameters of the beta distributed marginals are obtained through a beta regression. We use simple vine copula models in the form of Markov trees to describe the dependence structure. There is empirical evidence for different degrees of upper and lower tail dependence in the data. Therefore we compare different approaches using Gaussian, Gumbel, BB1- and BB7-copulas. Evaluation methods like the continuous ranked probability score (CRPS) and the variogram score (VS) are used to compare the predictive power of the various model approaches.

E1563: **Regularly varying random fields and local sequence alignment**
Presenter: **Hrvoje Planinic**, Faculty of Science, University of Zagreb, Croatia
Co-authors: Bojan Basrak

When considering ungapped local alignments of two independent i.i.d. sequences of letters from a finite alphabet, one usually constructs a (random) matrix which summarizes the local alignment scores. Under mild conditions, extreme values of this matrix appear in clusters along the diagonals, and it is known the number of such clusters is asymptotically Poisson distributed. This problem is motivated by applications in comparision of biological sequences. We discuss that, under suitable transformations, this problem can be analyzed using the theory of stationary regularly varying random fields, the key tool being the so–called tail process. In particular, using point processes we show that all extremes of the score matrix can asymptotically be approximated by a certain Poisson cluster process which is fully determined by the tail process.

E1925: **Bayesian semiparametric regression for heavy-tailed responses**
Presenter: **Junho Lee**, University of Edinburgh, United Kingdom
Co-authors: Miguel de Carvalho

Statistical modelling of extreme events-such as hurricane Dorian-is of the utmost importance for a variety of fields of research. We will propose a Bayesian semiparametric regression model for heavy-tailed responses. The proposed Bayesian smoothing method is motivated by the need of examining how the magnitude of extreme values may change along with a covariate. The model is built using a Pareto-type specification for the tail of the response, with covariates being modelled via generalized additive model (GAM) and Bayesian P-splines. Finally, details on computational implementation will be discussed over the talk, along with a set of illustrations on simulated and real data. We will cover the main results from a Monte Carlo experiment.

EO496 Room MAL 251 STATISTICAL METHODS FOR NON-EUCLIDEAN DATA Chair: Karthik Bharath

E0276: **Manifold valued data analysis of samples of networks, with applications in corpus linguistics**
Presenter: **Katie Severn**, University of Nottingham, United Kingdom
Co-authors: Ian Dryden, Simon Preston

Networks can be used to represent many systems such as text documents and brain activity, and it is of interest to develop statistical techniques to compare networks. A general framework is developed for extrinsic statistical analysis of samples of networks, motivated by networks representing text documents in corpus linguistics. Networks are identified by their graph Laplacian matrices, for which metrics, embeddings, tangent spaces, and a projection from Euclidean space to the space of graph Laplacians are defined. This framework provides a way of computing means, performing principal component analysis and regression, and performing hypothesis tests, such as for testing for equality of means between two samples of networks. The methodology is applied to the set of novels by Jane Austen and Charles Dickens.

E0629: **Mixed-effect model for the spatiotemporal analysis of heterogeneous longitudinal manifold-valued data**
Presenter: **Stephanie Allassonniere**, Paris Descartes University - INRIA, France
Co-authors: Vianney Debavelaere, Juliette Chevallier, Stanley Durrleman

A generic hierarchical spatiotemporal model for longitudinal manifold-valued data is presented. Data consist in repeated measurements over time for a group of individuals. We will consider heterogeneous populations. We will first introduce the single population model which allows us to estimate a group-average trajectory of evolution, considered as a piece-wise geodesic of a given Riemannian manifold. Individual trajectories of progression are obtained as random variations, which consist in parallel shifting and time reparametrization, of the average trajectory. These spatiotemporal transformations allow us to characterize changes in the direction and in the pace at which trajectories are followed. Then, the mixture case will be developed. We propose to estimate the parameters of the model using a stochastic version of the expectation-maximization (EM) algorithm, the Monte Carlo Markov Chain Stochastic Approximation EM (MCMC SAEM) algorithm with new tempering schemes. This generic spatiotemporal model is used to analyze the temporal progression of a family of biomarkers. This progression model estimates a normative scenario of the progressive impairments of several cognitive functions, considered here as biomarkers, during the course of Alzheimers disease. We also used this model to understand the response to antiangiogenic treatment in metastatic cancers.

E1266: **Detecting and correcting bias in phylogenetic tree inference**
Presenter: **Megan Owen**, Lehman College CUNY, United States

Phylogenetic tree inference is the problem of reconstructing the phylogenetic tree, which represents the evolutionary history of a set of organisms, from some genetic data, like DNA. Tree inference methods, like maximum likelihood and Bayesian MCMC, are known to exhibit bias, meaning they can produce trees with a tendency towards certain shapes or edge lengths. We propose a comprehensive method for detecting bias in tree shape and/or edge lengths using the Billera-Holmes-Vogtmann (BHV) tree space framework. The BHV tree space is a non-positively curved (or CAT(0)) geometric space containing all possible trees for a given set of leaves. The method employs a logarithm map to the tangent space, yielding a Euclidean space in which to perform analysis. We show that different tree inference methods have different biases and suggest methods for correcting them.

E1358: **The geometry of graph space: Towards graph-valued statistics**
Presenter: **Aasa Feragen**, Technical University of Denmark, Denmark
Co-authors: Anna Calissano, Simone Vantini

Graph-structured data are abundant in both nature and science, including blood- or lymphatic networks in the body; molecules in chemoinformatics; social networks in communication, or transportation networks in urban planning. While state-of-the-art machine learning such as neural networks are efficient for solving simple problems such as classification of graphs, which predicts a simple, discrete output (a class), they have a harder time solving problems whose answer is a graph, such as graph interpolation, decomposition of variance (PCA), or graph-valued regression. We discuss a space of attributed graphs with variable number of nodes, formed as a quotient of a space of adjacency graphs with respect to a node permutation group. The limitations and possibilities that derive from the geometric properties of the resulting graph-space will be explained. Moreover, problems with existing heuristics for statistics in this and related spaces will be discussed, and a novel strategy for computing statistics in quotient spaces will be presented. This novel strategy will be applied to the particular case of graph-space, defining principal component analysis as well as regression taking values in graph-space, along with heuristics that make them computationally practical.

E1932: **On the linear combination of chi-squares with applications to inference in shape and directional statistics**
Presenter: **Alfred Kume**, University of Kent, United Kingdom
Co-authors: Andrew Wood, Tomonari Sei

Some random matrix models adopted for statistical analysis of directions and shapes, rely on certain relationships with the linear combination of central and non-central chi-square random variables. Motivated initially from the directional statistics problems, we focus on the density function of such distributions and not on their cumulative distribution function which have been extensively covered in the literature. Our approach provides new insight by generating alternative characterisations of the relevant expressions for a range of distributions used in directional and shape inference. In addition, our results can be easily extended to some general expectations used for spike models used in random matrix theory. The expressions we obtain are more transparent for modelling purposes and with apparent stability in their numerical evaluation. We will illustrate our method with some examples.

EO745 Room MAL 252 OPTIMIZATION AND NEW STATISTICAL LEARNING TOOLS IN DATA SCIENCE Chair: Murat A Erdogdu

E1964: **Semi-orthogonal non-negative matrix factorization with an application in text mining**
Presenter: **Yutong Li**, University of Illinois at Urbana-Champaign, United States
Co-authors: Annie Qu, Ruoqing Zhu

Emergency Department (ED) crowding is a worldwide issue that affects the efficiency of hospital management and the quality of patient care. This occurs when the request for an admission ward-bed to receive a patient is delayed until an admission decision is made by a doctor. To reduce the overcrowding and waiting time of ED, we build a classifier to predict the disposition of patients using manually-typed nurse notes collected during triage. However, these triage notes involve high dimensional, noisy, and also sparse text data which makes model fitting and interpretation

difficult. To address this issue, we propose the semi-orthogonal non-negative matrix factorization (SONMF) for both continuous and binary design matrices to first bi-cluster the patients and words into a reduced number of topics. The subjects can then be interpreted as a non-subtractive linear combination of orthogonal basis topic vectors. These generated topic vectors provide the hospital with a direct understanding of the cause of admission. We show that by using a transformation of basis, the classification accuracy can be further increased compared to the conventional bag-of-words model and alternative matrix factorization approaches.

E1698: **Time-varying feature selection for longitudinal analysis**
Presenter: **Lan Xue**, Oregon State University, United States

Time-varying coefficient model selection and estimation is proposed based on the spline approach, which is capable of capturing time-dependent covariate effects. The new penalty function utilizes local-region information for varying coefficient estimation, in contrast to the traditional model selection approach focusing on the entire region. The proposed method is extremely useful when the signals associated with relevant predictors are time-dependent, and detecting relevant covariate effects in the local region is more scientifically relevant than those of the entire region. Our simulation studies indicate that the proposed model selection incorporating local features outperforms the global feature model selection approaches. The proposed method is also illustrated through a longitudinal growth and health study from National Heart, Lung and Blood Institute.

E0418: **Bandit algorithms for nonstationary online nonconvex optimization**
Presenter: **Krishnakumar Balasubramanian**, University of California, Davis, United States

Bandit algorithms have been predominantly analyzed in the convex setting with function value based stationary regret as the performance measure. We propose and analyze bandit algorithms for nonconvex problems with nonstationary regret as the performance measure. Specifically, we consider nonstationary regret measures in terms of function-values, first-order and second-order stationary solutions. We provide regret bounds for the Gaussian Steins identity based bandit algorithm for two classes of nonconvex functions, in both the low and high-dimensional settings, in terms of function-value and first-order stationary solutions based regret measures. We also propose online and bandit versions of the cubic regularized Newtons method. The bandit version is based on estimating the Hessian matrices in the bandit setting, based on second-order Gaussian Steins identity. We provide nonstationary regret bounds in terms of second-order stationary solutions, that have interesting consequences for avoiding saddle points in the bandit setting.

E1357: **Global nonconvex optimization with discretized diffusions**
Presenter: **Murat A Erdogdu**, University of Toronto, Canada

An Euler discretization of the Langevin diffusion is known to converge to the global minimizers of certain convex and non-convex optimization problems. We show that this property holds for any suitably smooth diffusion and that different diffusions are suitable for optimizing different classes of convex and non-convex functions. This allows us to design diffusions suitable for globally optimizing convex and non-convex functions not covered by the existing Langevin theory. The non-asymptotic analysis delivers computable optimization and integration error bounds based on easily accessed properties of the objective and chosen diffusion. Central to our approach are new explicit Stein factor bounds on the solutions of Poisson equations. We complement these results with improved optimization guarantees for targets other than the standard Gibbs measure.

E1441: **Convergence rate of block-coordinate maximization Burer-Monteiro method for solving large SDPs**
Presenter: **Nuri Vanli**, MIT, United States
Co-authors: Murat A Erdogdu, Pablo Parrilo, Asuman Ozdaglar

Semidefinite programming (SDP) with diagonal constraints arise in many problems, such as Max-Cut, community detection and group synchronization. Although SDPs can be solved to arbitrary precision in polynomial time, generic convex solvers do not scale well with the dimension of the problem. In order to address this issue, it has been previously proposed to reduce the dimension of the problem by appealing to a low-rank factorization, and solve the subsequent non-convex problem instead. We present coordinate ascent based methods to solve this non-convex problem with provable convergence guarantees. In particular, we prove that the block-coordinate maximization algorithm applied to the non-convex Burer-Monteiro approach globally converges to a first-order stationary point with a sublinear rate. We also show that the block-coordinate maximization algorithm is locally linearly convergent to a local maximum under local weak strong concavity assumption. We establish that this assumption generically holds when the rank of the factorization is sufficiently large. Furthermore, we propose an algorithm based on the block-coordinate maximization and Lanczos methods that is guaranteed to return a solution that provide $1 - O(1/r)$ approximation to the original SDP, where r is the rank of the factorization, and we quantify the number of iterations to obtain such a solution. This approximation ratio is known to be optimal under the assumption of the unique games conjecture.

EO699 Room MAL 253 RECENT ADVANCES ON ROC CURVES ESTIMATION Chair: Graciela Boente

E0216: **Improving the biomarker diagnostic capacity via functional transformations**
Presenter: **Pablo Martinez-Camblor**, Geisel School of Medicine Dartmouth College, United States

The use of the area under the receiver-operating characteristic, ROC, curve (AUC) as an index of diagnostic accuracy is overwhelming in fields such as biomedical science and machine learning. A larger AUC has become synonymous with a better performance. A functional transformation of the marker values has been proposed for increasing the AUC and then the diagnostic accuracy. The classification process is based on some regions which support the decision made; one subject is classified as positive if its marker is within this region, and as negative otherwise. We study the capacity of improving the classification performance of univariate markers via functional transformations and the impact of this transformation on the final classification regions based on a real-world dataset. Particularly, we consider the problem of determining the gender of a subject based on the Mode frequency of his/her voice. The shape of the cumulative distribution function of this characteristic in both the male and female groups makes the classification problem useful for illustrating the differences between having valuable diagnostic rules and obtaining an optimal AUC. Our point is that improving the AUC by means of a functional transformation can produce classification regions with no practical interpretability. We propose to improve the classification accuracy by making the selection of the classification subsets more flexible while preserving their interpretability.

E0800: **An approach to evaluate and compare biomarkers to diagnose a disease**
Presenter: **Maria del Carmen Pardo**, Complutense University of Madrid, Spain
Co-authors: Christos T Nakas, Alba Franco-Pereira

The index of the Area Under the ROC curve (AUC) reflects the amount of separation of the biomarker distributions in the two samples of subjects derived from the non-diseased and diseased populations. Along with the AUC, the maximum of the Youden index, J, is often used for the comparison of competing biomarkers. We study the utility of the Length of the binormal model-based ROC Curve (LoC) as an index of diagnostic accuracy for biomarker evaluation. In a simulation study, the performance of LoC is compared with approaches based on AUC and J, both for the case of the assessment of a single biomarker and for the comparison of two biomarkers, in a parametric framework. We provide an interpretation for the proposed index and illustrate with an application on biomarkers from a colorectal cancer study.

E0303: **Robust estimation of ROC curves with covariates**
Presenter: **Ana Maria Bianco**, Universidad de Santiago de Compostela, Depto. de Estadistica e Inv. Operativa, Spain
Co-authors: Graciela Boente, Wenceslao Gonzalez-Manteiga

Receiver Operating Characteristic (ROC) curves are a useful graphical tool to measure the discriminating power of a continuous variable, such as diagnostic variable or a marker. They are employed to quantify the accuracy of the marker to distinguish between two conditions or classes. As with any classifier, the assignations are not perfect and may lead to classification errors. In practical situations, the discriminatory effectiveness of the marker under study may be affected by several factors. When for each individual additional information is available, it is sensible to include it in the ROC analysis. The aim is to show the instability of the conditional ROC curve in presence of outliers and also to provide robust estimators when covariates are available. A semiparametric approach is followed, where robust parametric estimators are combined with weighted empirical distribution estimators based on an adaptive procedure that downweights outliers. The consistency of the proposal is discussed. Through a Monte Carlo study, the performance of the proposed estimators is compared with that of the classical ones in clean and contaminated samples.

E0590: **Robust inference for the covariate-specific ROC curve and its associated summary indices**
Presenter: **Vanda Lourenco**, Faculty of Sciences and Technology - New University of Lisbon, Portugal
Co-authors: Miguel de Carvalho, Vanda Inacio

Accurate diagnosis of disease is of critical importance in health care and medical research. The receiver operating characteristic (ROC) curve is the most popular tool for evaluating the discriminatory ability of continuous biomarkers. In practice, the performance of a test/biomarker can depend on covariates (e.g., age and/or gender). In order to take covariate information into account, the covariate-specific ROC curve has been proposed as a way of evaluating how the accuracy of the biomarker changes as a function of such covariates. We develop a robust and flexible model for conducting inference about the covariate-specific ROC curve and its associated covariate-specific summary indices, that safeguards against atypical biomarker observations while accommodating for nonlinear covariates effects. Specifically, we postulate a location-scale regression model for the test outcomes in each group, combining additive B-splines regression and M-estimation for the mean function with the residuals being estimated via a weighted empirical distribution function. Simulation results show that our approach successfully recovers the true covariate-specific ROC curve and corresponding summary indices on a variety of data contamination scenarios. The adequacy of the method is further illustrated using data on age-specific accuracy of glucose as a biomarker of diabetes.

E1176: **Comparing ROC curves conditioned to a multidimensional covariate**
Presenter: **Aris Fanjul Hevia**, Universidad de Santiago de Compostela, Spain
Co-authors: Wenceslao Gonzalez-Manteiga, Juan-Carlos Pardo-Fernandez, Ingrid Van Keilegom

The comparison of Receiver Operating Characteristic (ROC) curves is a commonly accepted way of comparing the discriminatory capability of different diagnostic variables of a certain disease. There are several methodologies discussed in the literature for making that comparison, although most of them do not consider the possible effect that the presence of covariates can have in the performance of the test. A new test is proposed for comparing ROC curves conditioned to the value of a multidimensional covariate. Projections will be used for transforming the problem into a one dimensional approach easier to handle. Some simulations of the new methodology and an illustration on a real dataset application are presented.

EO855 Room SH349 RECENT ADVANCES IN GENOMIC PREDICTION Chair: Reka Howard

E0353: **The performance of genotype-to-phenotype models accounting for large-effect loci, epistasis, and pleiotropy**
Presenter: **Alexander Lipka**, University of Illinois, United States

Models that reflect the multifaceted contributions of genomic loci have a potential to facilitate unprecedented quantification of the genetic architecture underlying various traits and increase genomic selection (GS) prediction accuracies. The performance of statistical models is evaluated for traits with contrasting genetic architectures. These traits were simulated using marker data in maize, sorghum, and humans. These simulation studies revealed that including peak-associated markers from a genome-wide association study (GWAS) of a training set as fixed-effect covariates in an RR-BLUP genomic selection model is capable of decreasing prediction accuracy, increasing the variability of prediction accuracy across replicate traits, and increasing the bias of predictions compared to a standard RR-BLUP GS model. The studies also suggest that a model quantifying the simultaneous contribution of additive and two-way epistatic loci is capable of identifying and distinguishing between simulated additive and epistatic quantitative trait nucleotides (QTNs). Finally, the latest results from an ongoing simulation study seeking to explore the ability of a multi-trait GWAS model to identify simulated pleiotropic QTN is presented. These results will underscore current efforts to refine GS models that that go beyond univariate models accounting for only additive marker effects.

E0459: **Predicting environmental response of crop plants via the integration of models from different disciplines**
Presenter: **Hiroyoshi Iwata**, The University of Tokyo, Japan

Prediction of crop responses to environments is essential to efficiently develop new varieties that can adapt to the target environment. Genetic ability of crop plants can be predicted based on genome-wide DNA polymorphisms (i.e., genomic prediction). The environmental response of crop plants, however, is difficult to predict because it is influenced not only by genetic but also by environmental factors. To predict the environmental responses, it is necessary to extend genomic prediction to take into account the variations caused by environmental factors in the prediction. The integration of models from different disciplines will be necessary for the extension of genomic prediction. Crop simulation models, which enable the prediction of crop growth under given environments, and machine learning models associating the crop response to environments with environmental factors are good candidates for the integration. We will introduce methods for integrating genomic prediction models with crop models and machine learning models to predict the environmental responses of crop plants. We will also introduce methods for integrating multi-omics data to predict the crop responses to environments.

E0461: **Response surface analysis of genomic prediction accuracy values using quality control covariates in soybean**
Presenter: **Reka Howard**, University of Nebraska - Lincoln, United States

An important tool for selection purposes and to increase yield in plant breeding is genomic prediction. Genomic prediction is a technique where molecular marker information and phenotypic data are used to predict the phenotype of individuals for which only marker data are available. Higher prediction accuracy can be achieved not only by using efficient models but also by using quality molecular data. The steps of a typical quality control of marker data include the elimination of markers with certain level of minor allele frequency (MAF) and missing marker values and the imputation of missing marker values. We evaluated how the prediction accuracy is influenced by the combination of 12 MAF values, 27 different percentages of missing marker values, and 2 imputation techniques. We constructed a response surface of prediction accuracy values as a function of MAF and percentage of missing marker values using soybean data. We found that both the genetic architecture of the trait and the imputation technique affect the prediction accuracy. For the corresponding combinations MAF-percentage of missing values we observed that implementing the random forest imputation increased the number of markers by 2 to 5 times than the simple nave imputation method that is based on the mean allele dosage of the non-missing values at each loci. There is not a unique strategy (combination of the QCs and imputation method) that outperforms the results of the others for all traits.

E0414: **Genome and phenome statistical models and methods for prediction**
Presenter: **Jose Crossa**, University of Nebraska-Lincoln, United States
Co-authors: Osval Montesinos-Lopez, Paulino Perez-Rodriguez, Reka Howard, Jaime Cuevas, Diego Jarquin, Abelardo Montesins-Lopez

In the last years genome and phenome models have been developed for the prediction of unobserved individuals using dense molecular markers and high throughput phenotype (HTP) informations. Statistical models include single and multi-traits, and single and multi-environments as well as several HTP and near infrared spectroscopy (NIR) information with the objective of increase the prediction accuracy of grain yield, and other

traits on unobserved individuals. Increase in prediction accuracy over the genomic best linear unbiased predictor (GBLUP) was achieved by means of the Gaussian kernel model including genomic by environments interaction (GE). A Bayesian multi-trait multi-environment model can efficiently exploit correlated traits and environments and thus increasing the prediction accuracy of about 10% over the single trait, single environment model. Bayesian models requires intense computational resources. Deep Machine Learner (DL) models with densely connected network architecture were developed with the objective of using less computing resources and accommodate extensive data sets. Under certain circumstances DL were competitive with other well established models. Although implementing the multi-trait DL models is feasible and practical in the genomic prediction context it is challenging due to the large number of hyper-parameters involved. Deep Kernel method has been studied and results compared with other kernel methods.

E1122: **Models for autotetraploid genomics in the context of GWAS and GS**
Presenter: **Patricio Munoz Del Valle**, University of Florida, United States

Autopolyploidy occurs in many important crops (e.g. alfalfa, blueberry, potato). Autopolyploidy can significantly reduce the speed at which plant breeding can concentrate favorable alleles. This is because polyploid genomes and specifically autopolyploid genomes are more complex than diploids. Polysomic inheritance, higher allele dosage, double reduction and higher probability of non-additive effects are some of these complexities. Nevertheless, with the advances in technology and more affordable prices, more autopolyploid breeding programs are starting to genotype their populations to either find genes associated to important traits or for applying genomic selection (GS). However, whether to account for dosage or non-additive effects is still a matter of discussion in autopolyploids. We will illustrate, using a population of blueberry as a model, the effect that dosage can have in the capacity to discover quantitative trait loci in a genome-wide association analysis (GWAS) as well as the effect on the prediction ability on genomic selection models. Also, the effect of accounting for dominance in the context GS and the current challenges to apply genomics in polyploid breeding will be discussed.

EC811 Room MAL G14 CONTRIBUTIONS IN STOCHASTIC PROCESSES Chair: Serguei Dachian

E1602: **Bootstrap method for misspecified ergodic stochastic differential equation models**
Presenter: **Yuma Uehara**, The Institute of Statistical Mathematics, Japan

Ergodic stochastic differential equation models driven by Levy processes under model misspecification are considered. A Gaussian quasi-likelihood based method serves as a good device to estimate drift and scale parameters in the models. However the correction of misspecification bias makes it difficult to construct a consistent estimator of the asymptotic variance of the Gaussian quasi-likelihood estimator, and thus confidence intervals and hypothesis testing. For such a problem, we propose a (blocking) weighted bootstrap method to directly approximate the asymptotic distribution of the estimator. We show that the approximation theoretically works well, and present some numerical experiments.

E1611: **Data driven time scale for diffusion processes in YUIMA**
Presenter: **Shoichi Eguchi**, Osaka University, Japan

Parametric estimation of diffusion processes with unknown sampling stepsize has been studied, and estimators of model parameters and sampling stepsize has been constructed in a fully explicit way. Based on that, we create the function which can estimate model parameter and sampling stepsize for ergodic diffusion processes in R package yuima. We will first overview the estimation method of model parameters and sampling stepsize and then explain the specification of the created function. Some numerical examples are given in order to show how to use the function.

E1878: **A stochastic extension of the T growth model**
Presenter: **Antonio Barrera**, Universidad de Malaga, Spain
Co-authors: Patricia Roman-Roman, Francisco Torres-Ruiz

Stochastic growth models are of great importance due to the ability to include random influences produced by internal and external conditions. Usually, these models are extensions of deterministic models related with growth curves such as the classical logistic, Weibull or, more recently, hyperbolastic. One of the most interesting recent curves is the one associated with the T model, which has become useful in modelling bacterial growth or proliferation and regression of cancer cells. This deterministic model is able to represent sigmoidal and biphasic growth with great accuracy, but it does not take into account random effects. A stochastic extension based on the T model is proposed. Starting from a parametric modification of the original curve, a time nonhomogeneous diffusion process is built providing the same mean behaviour. The main issue is related with the estimation of the parameters, because of the complexity of the model. In order to deal with this problem, different strategies based on the reformulation of maximum likelihood equations and the use of metaheuristic algorithms, are considered. Finally, usefulness of the stochastic model is illustrated by performing some practical applications.

E1433: **Rate of convergence to alpha stable law using Zolotarev distance**
Presenter: **Solym Manou-Abi**, CUFR de Mayotte and IMAG-Montpellier, France

The question of the rate of convergence to stable laws is considered using arguments based on the Zolotarev distance to prove bounds. We provide a rate of convergence to stable random variable in the generalized CLT, that is, for the partial sums of independent identically distributed random variables which are not assumed to be square integrable.

E1599: **Convergence of the logarithm of empirical characteristic function in stable laws**
Presenter: **Annika Krutto**, University of Tartu, Estonia

The properties of empirical characteristic function are well studied. However, in estimating the parameters of stable laws the logarithm of empirical characteristic function, which has not much been studied, may be more useful. The flexible 4-parameter stable laws can capture the fuzzy dynamics and large fluctuations that result from stochastic processes occurring in diverse fields of finance, insurance, and climatology. The estimation of the parameters of stable laws is complicated due to the fact that many have infinite moments and, with a few exceptions, the densities cannot be explicitly expressed in the terms of elementary functions, causing many practitioners to avoid stable laws. The convergence rate of the real part of the logarithm of the empirical characteristic function in stable laws is studied for fixed sample size n. The results can give significant contribution in solving the problem of optimal argument selection in the simple empirical characteristic function based estimation of stable laws.

EC808 Room MAL 254 CONTRIBUTIONS IN FUNCTIONAL DATA ANALYSIS Chair: Simone Vantini

E1782: **Exponential smoothing with locality parameter for functional data in robust forecasting of economic phenomena**
Presenter: **Przemyslaw Jasko**, Cracow University of Economics, Poland
Co-authors: Daniel Kosiorowski, Jerzy Rydlewski

A variety of economic phenomena may be described as random functional variable or as a functional time series, i.e., as a family of such variables indexed by time. A functional view on the phenomena is very often more natural than classical view as in the functional view we do not need to divide economic system into separate parts. Economic functional time series very often consist of functional outliers of various kinds. A development of effective robust methods of forecasting of the economic time series is an issue of a prime importance for theory of economics as well as for applied economics and empirical finance. Due to a temporal dependency between observations well-know strategies of dealing with contaminated datasets are not directly applicable in case of the functional time series. We propose a version of exponential smoothing predictor for functional data based on selected depths for functional data. We discuss its properties and compare it with other predictors known form the

literature. Theoretical considerations are illustrated via results of simulation studies and empirical example related to an analysis of a process a development of a city based on satellite maps.

E1848: **Impact point selection in semiparametric multi-functional regression**
Presenter: **Silvia Novo**, University of A Coruna, Spain
Co-authors: German Aneiros, Philippe Vieu

A new sparse regression model is proposed in the functional data context, which incorporates the influence of two functional random variables in a scalar response: one of them is included linearly, but trough the high-dimensional vector formed by its discretized observations and, the other one, trough a single-index structure. For this semiparametric model, two new algorithms for selecting impact points in the linear part and for estimating the model are presented. Both procedures take advantage of the functional origin of the linear covariates. Some asymptotic results will support theoretically both methods. Finite sample experiments and a real data application will ensure their good practical behaviour.

E0313: **Dependent random measures indexed by a functional covariate**
Presenter: **Emmanuel Bernieri**, University of Edinburgh, United Kingdom
Co-authors: Miguel de Carvalho

The analysis of functional data is explored in a nonparametric Bayesian context. Specifically, we devise priors in the space of all conditional distributions, for the setting where the interest is on conditioning on a sophisticated object such as a random function. The proposed model can be regarded as an infinite mixture of functional linear regression models. A specific version of the proposed model is explored in detail, which consists of a Dependent Dirichlet Process (DDP) whose regression functions include inner products between a functional covariate and coefficient function. We illustrate the proposed methods using simulated and real data, and evaluate the accuracy of the methods through a simulation study.

E1854: **Approaches for extending multiple imputation to handle scalar and functional data**
Presenter: **Adam Ciarleglio**, George Washington University, United States

Missing data are a common problem in biomedical research. Valid approaches for addressing this problem have been proposed and are regularly implemented in applications where the data are exclusively scalar-valued. However, with advances in technology and data storage, biomedical studies are beginning to collect both scalar and functional data, both of which may be subject to missingness. We propose extensions of multiple imputation with predictive mean matching and imputation by local residual draws as two approaches for handling missing scalar and functional data. The two methods are compared via a simulation study and applied to data from a study of subjects with major depressive disorder for which both clinical (scalar) and imaging (functional) data are available.

E2013: **Functional data analysis application in TEC disturbance caused by Tsunami**
Presenter: **Ryuichi Kanai**, University College London, United Kingdom
Co-authors: Serge Guillas

The ionospheric plasma disturbance caused by a large tsunami after subduction earthquake can be detected by measurement of the total electron content (TEC) between global positioning system (GPS) satellites and their receivers. TEC depression which is termed as tsunami ionospheric hole (TIH) lasting for several tens of minutes is formed above the tsunami source area. However, even if the network of GPS receiving stations is dense, it is impossible to detect all data around tsunami source. Firstly, we propose a method using bivariate spline fitting to interpolate the detected TEC data. By this method, it becomes possible to estimate tsunami source using this fitted data. In addition, using this fitting method, less GPS receivers can show a sufficient estimation of TEC disturbance. Secondly, with this fitted data, we succeeded in evaluating the TEC dip quantitatively using Functional Principal Component Analysis (FPCA) score on a fixed longitude. Based on the FPCA scores, we can estimate when the TEC disturbance becomes the largest. Thirdly, applying the functional linear regression to these data, we can estimate the TEC disturbance data at least 30 seconds before.

EC801 Room MAL 355 CONTRIBUTIONS IN APPLIED STATISTICS Chair: Claudia Neves

E1856: **Statistical inference in hydraulic tomography with wavelet-based priors**
Presenter: **Philipp Wacker**, FAU Erlangen-Nuernberg, Germany
Co-authors: Peter Knabner

Hydraulic tomography is a technique for inferring subsurface properties like hydraulic permeability as a function of the domain. Probes are alternatingly used as pumps and pressure measuring device. The process can be compared to medical imaging technology (only that we are interested in the subsurface's inner composition, not some patient's). This constitutes an infinite-dimensional statistical inverse problem and is usually solved by a geostatistical approach. We will try to tackle this problem by setting a prior distribution which is governed by random superpositions of wavelet functions (similar to more common priors which are random superpositions of sine and cosine functions).

E1889: **Clustering time-varying quaternion data: Application to the detection of gait abnormalities**
Presenter: **Pierre Drouin**, Laboratoire de mathematiques Jean Leray - Nantes / UmanIT, France
Co-authors: Aymeric Stamm, Lise Bellanger, Laurent Chevreuil, Vincent Graillot

Wearable motion sensors have become more and more compact and affordable. As such, they might become useful to the neurologist for achieving objective and quantitative assessment of walking disability in patients diagnosed with a neurodegenerative disorder, such as Multiple Sclerosis. The company UmanIT and the Department of Mathematics Jean Leray in Nantes have developped a solution for facilitating gait analysis by collecting data using an inertial measurement unit embedded in a motion sensor. The device is clipped on the belt and records hip rotation over time as a sequence of quaternions. After pre-processing, a set of walking cycles (i.e. all movements made between two equivalent positions of a given foot) is provided by the device. To test the hypothesis that differences between hip motion during walking cycles are related to walking disability, an experiment was conducted on healthy volunteers who performed a walking test under two conditions: (i) free motion and (ii) wearing a knee blocking splint. Clustering methods will be adapted to account for the non-Euclidean nature of quaternion data. In particular, approaches borrowed from time series (e.g. dynamic time warping) and from functional data (e.g. k-mean alignment) will be extended. As part of a benchmarking study, these methods will be evaluated based on their performance to create two groups that separate walking cycles according to the two test conditions and compared to traditional clustering methods.

E1904: **Assessing the representativeness of non-probability surveys: The case of public library users' survey in Florence**
Presenter: **Emilia Rocco**, University of Florence, Italy
Co-authors: Chiara Bocci, Alessandra Petrucci

With the growing spread of big data, as a potential source of low-cost and timely data, there is a greater interest in whether and how it is possible to use data from non-probability samples to make inference. The main problem when the data generating mechanism is unknown and/or is presumably very different from random sampling is that estimators of population characteristics must be assumed to be biased, unless convincing evidence to the contrary is provided. Therefore, it becomes necessary to identify indicators/measures of the potential risk of non-random selection bias and to adopt predictive inference methods able to remove this bias. Users registered in the public library system of the city of Florence are a non-random subset of the resident population. Among these, the respondents to a web survey for the evaluation of the libraries' services are the result of a further process of self-selection. The representativeness of the respondents' data is investigated in order to evaluate the possibility to make inference on the whole Florentine population.

E1691: **Substitution bias of the consumer price index in Poland**
Presenter: **Aleksandra Halka**, Narodowy Bank Polski, Poland
Co-authors: Agnieszka Leszczynska-Paczesna
The focus is on the problem of bias in the measure of inflation as provided by the price index of consumer goods and services (CPI) in Poland. We estimate the size of the bias resulting from substitution effect and the application of plutocratic weights in index calculation. The results show a downward substitution bias of Polish CPI index which is rather unusual, and may stem form frequent adjustments in the weights used for CPI calculation as well as to a faster-than-CPI rise in the prices of those goods and services the demand for which is relatively inelastic. It was found that the CPI (plutocratic) index for Poland was lower than the democratic one.

E0620: **Linking climate and dengue using the integrated nested Laplace approximation and the SPDE approach**
Presenter: **Stephen Jun Villejo**, University of the Philippines, Philippines
The primary goal is to perform a combined analysis of two spatially misaligned data using the stochastic partial differential equations approach (SPDE) estimated using the Integrated Nested Laplace Approximation (INLA) method. In particular, the two datasets considered are measurements of several climate indicators from several weather stations and the incidence of dengue by province in the Philippines. The former data is point-level while the latter is area-level. A continuously-indexed Gaussian process for the climate variables is assumed and is approximated by a discretely-indexed Gaussian Field (GF) via triangulation with the additional Gaussian Markov Random Field assumption. The obtained estimate of the GF is projected on blocks, which are subsets of the 2D space, corresponding to the provincial boundaries in the entire domain area. The Besag-York-Mollie (BYM) specification is used in providing a link between the incidence of dengue and the projections of the Gaussian process for all the climate variables. The results show a significant association between relative humidity and temperature with the incidence of dengue. Moreover, maps of the relative risks and excess risks were also computed using the INLA and Markov Chain Monte Carlo (MCMC) methods with the BYM model as the baseline model. The MCMC method had some convergence issues, but the estimates of the relative risks and excess risks from the two methods are almost the same.

CO743 Room Bloomsbury COMMODITY MARKETS Chair: Ana-Maria Fuertes

C0768: **Hazard fear in commodity markets**
Presenter: **Ana-Maria Fuertes**, Cass Business School - City University London, United Kingdom
Co-authors: Adrian Fernandez-Perez, Joelle Miffre, Marcos Gonzalez-Fernandez
The aim is to introduce a commodity futures return predictor related to fear about impending weather, disease, geopolitical and economic hazards that can shift the commodity supply or demand. Exploiting the commodity hazard-fear characteristic as a trading signal in a long-short portfolio framework, we find a sizeable and significant commodity premium. The hazard-fear premium reflects compensation for known factors such as basis, momentum and illiquidity risks, but is not subsumed by them. Exposure to hazard-fear is strongly priced in the cross-section of individual commodity futures returns and commodity portfolios beyond known risk factors. We identify a strong role for general investor sentiment in the commodity hazard-fear premium.

C0879: **Crudeo oil return predictability revisited**
Presenter: **Thomas Conlon**, University College Dublin, Ireland
Out-of-sample crude oil return predictability is re-examine by using a large set of predictors and finds that previous studies overstate the evidence of predictability. This overstatement comes from the significantly high first-order autocorrelation in monthly average crude oil returns, computed from monthly averages of daily prices of West Texas Intermediate crude oil spot provided by the U.S. Energy Information Administration. Monthly average returns would not have been available to the forecaster in real time. Following the convention in the stock, bond, currency, and commodity return predictability literature, we compute crude oil returns using end-of-month prices. Using this return series, we find no evidence of crude oil return predictability, reversing the conclusion of previous studies. More specifically, individual predictive and combination model forecasts of crude oil returns generated using popular economic and technical indicator variables fail to outperform the simple historical average return forecast.

C1367: **Long-run reversal in commodity returns: Insights from seven centuries of evidence**
Presenter: **Adam Zaremba**, University of Dubai, United Arab Emirates
Co-authors: Adam Zaremba, Robert Bianchi, Mateusz Mikutowski
The longest study of long-run reversal in commodity returns ever conducted is performed. Using a unique dataset of prices of 52 agricultural, industrial, and energy commodities, we examine the price behaviour for the years 1265 to 2017. The findings reveal a strong and robust long-run reversal effect. The returns of the past one to three years negatively predict subsequent performance in the cross-section of returns. The long-run reversal effect is present in both agricultural and non-agricultural commodity returns across all centuries and is independent of market states. The long-run reversal cannot be explained by macroeconomic risks. The phenomenon is elevated in more volatile commodities and in periods of high return dispersion.

C1498: **Media tone in commodity markets**
Presenter: **Nan Zhao**, Cass Business School, City, University of London, United Kingdom
Co-authors: Ana-Maria Fuertes
Media tone measures, as a proxy for sentiment, are constructed for a cross-section of commodities in the main sectors-energy, agriculture, and metals-based on the aggregate textual tone retrieved from news in Bloomberg terminals. Building on the long-short portfolio with equal-weighted constituents sorted on their hedging pressure characteristic as a baseline, we design an alternative hedging pressure sentiment-weighted portfolio that weighs the constituents according to their media tone over the preceding week. A constituent of the long (short) leg of the baseline portfolio is assigned a smaller (larger) weight if its media tone is net positive (negative), since this favourable (adverse) sentiment is likely to be already reflected in the current futures price pushing it upwards (downwards). Thus, the expected return would be lower (higher) than it would have been anticipated according to the hedging pressure characteristic alone. We find that the media tone corrected portfolio could earn a significantly higher return than the benchmark. Then, we construct a media tone-based sentiment index per commodity. Using it as a trading signal in an out-of-sample portfolio analysis, we find a sizeable and significant sentiment premium after controlling for traditional commodity risk factors. The findings suggest that sentiment can predict the cross-section commodity futures return.

C1547: **On the existence of a risk free asset: The first empirical test of zero-beta CAPM using T-bills and gold**
Presenter: **Zhen He**, University of York, United Kingdom
Co-authors: Jacco Thijssen, Fergal OConnor
The aim is to test by the first time whether government Treasury bill (T-bill) or gold can be a proper risk-free asset in the zero-beta Capital Asset Pricing Model (CAPM). We start with the CAPM to test gold as a zero-beta asset. However, the results of statistical power of the CAPM regression show that CAPM is not consistently a sufficient method. Then we follow the hypothesis of the zero-beta asset in the zero-beta CAPM. Wald test and Likelihood ratio test are applied in the zero-beta CAPM to investigate whether T-bill or gold is a zero-beta asset. This is the first empirical test of zero-beta asset in zero-beta CAPM. This approach allows the data to examine if T-bill or gold is a zero-beta asset, rather than arbitrarily setting T-bill as the zero-beta asset and further as the risk-free asset. Due to the results of Wald test in zero-beta CAPM, we find that neither the T-bill nor

gold is a zero-beta asset, let alone risk-free asset. This provides the evidence that there is no existence of the risk-free asset, and that the assumption of classic CAPM does not hold.

CO683 Room G11 BAYESIAN FINANCIAL ECONOMETRICS Chair: Jia Liu

C1310: **Bayesian nonparametric methods for analysing macroeconomic time series**
Presenter: **Maria Kalli**, University of Kent, United Kingdom

The analysis of macroeconomic time series often involves the use of a vector autoregressive (VAR) model. VAR models provide a framework for the analysis of the complex joint dynamics present between macroeconomic series, but they have been criticised for their unrealistic assumptions (linearity, homoscedasticity, Gaussianity). We are going to describe how Bayesian non-parametric methods can be used to directly model the stationary and transition densities of such a multivariate system. This approach allows for nonlinearity in the conditional mean, heteroscedasticity in the conditional variance, and non-Gaussian innovations. It can also allow for non-stationary. Our empirical applications lie within the study of monetary policy and macro financial linkages within the aggregate economy. We find that the Bayesian nonparametric VAR (BayesNP-VAR) model predictively outperforms competing models.

C1277: **A dynamic Bayesian nonparametric model**
Presenter: **John Maheu**, McMaster University, Canada
Co-authors: Yong Song

A Bayesian nonparametric approach is designed to model dynamic changes in an unknown distribution through time. The distribution changes can be interpreted as structure change and we discuss how to perform inference on break dates. A new efficient MCMC routine is provided to estimate the model. Applications to housing data and bank data with comparison to other nonparametric models show the model to work well.

C0375: **A Bayesian nonparametric approach on model combination for short-term interest rates**
Presenter: **Qiao Yang**, ShanghaiTech University, China
Co-authors: John Maheu

The dynamics of short-term interest rates are important input into pricing models of the term structure of interest rates. Most of the works have focused on adding extra components into the model for improving forecast accuracy. We take another direction which uses the model combination framework to improves the forecast performance. We introduce Bayesian nonparametric approach to extend a previous one from fixed number states into infinite dimension and combine popular discrete time short-rate models. The new approach shows significant improvement in density forecasts to existing approaches and strong evidence of interest rated model dynamics are documented.

C1037: **Do better return density forecasts lead to economic gains in portfolio allocation?**
Presenter: **Chenxing Li**, McMaster University, Canada

This paper investigates the relationship between statistical improvements in density forecasts of returns and actual economic gains in portfolio allocation for a risk-averse investor. To aid this investigation, this paper proposes a new multivariate Bayesian semiparametric model that has better out-of-sample density forecasts than benchmark models. Results show that this more sophisticated econometric model does provide positive economic gains whether the investors utility is CRRA, CARA or quadratic. The economic gain diminishes when the investor is more risk-averse because she is moving away from risky investment positions.

C0593: **Bayesian nonparametric covariance estimation with noisy and nonsynchronous asset prices**
Presenter: **Jia Liu**, Saint Mary's University, Canada

A Bayesian nonparametric approach is proposed to estimate the ex-post covariance matrix of asset returns from high-frequency data in the presence of market micro-structure noise and non-synchronous trading. Several contributions are made. First, pooling is used to group returns with similar covariance matrices to improve estimation accuracy. Second, a new synchronization method of observations based on data augmentation is introduced. Third, the estimator is guaranteed to be positive definite. Finally, the new approach delivers exact finite sample inference without relying on asymptotic assumptions. All of those benefits lead to a more accurate estimator, which is confirmed by Monte Carlo simulation results. In real data applications, the proposed covariance estimator results in better portfolio choice outcomes.

CO210 Room G3 FINANCIAL MODELING AND STATISTICS Chair: Jan Vecer

C1809: **Utility-based model selection and model averaging**
Presenter: **Jan Vecer**, Charles University, MFF, Czech Republic
Co-authors: Robert Navratil

A novel approach is presented to model selection and model averaging based on economic theory. Model prediction is studied in the form of a distributional opinion about a random variable X. We show how to test this prediction against alternative views. Different model opinions can be traded on a hypothetical market that trades their differences. Using a utility maximization technique, we describe such a market for any general random variable X and any utility function U. We specify the optimal behavior of agents and the total market that aggregates all available opinions and show that a correct distributional opinion realizes profit in expectation against any other opinion, giving a novel technique for model selection. Analytical solutions are available for random variables from the exponential family. We determine the distribution corresponding to the aggregated view of all available opinions, giving a novel technique for model averaging.

C1835: **Long term portfolio protection**
Presenter: **Robert Navratil**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Jan Vecer

The purpose is to discuss novel approaches how to protect potential portfolio losses on long term horizons in the scale of several decades, which is a typical investment horizon of pension fund investments. The prices of existing financial products, such as put or call options, are increasing as a function of maturity, and their prices quickly take a significant percentage of the underlying assets. In this respect, such financial products become prohibitively expensive on horizons longer than a couple of years at most. In addition, these contracts tend to insure only static rather than actively traded portfolios which are more appropriate for pension funds. Thus it is desirable to have a protection of actively traded portfolio, where the client is free to move her wealth within different asset classes, while the portfolio value is protected against any trading losses. This is a generalization of a previously studied contract known as a passport option, but in our setup, the price of this contract is small enough to be attractive on 20-30 year investment horizons and thus the respective hedging strategy can be potentially embedded in pension fund products.

C1938: **Social security benefit valuation, risk, and optimal retirement**
Presenter: **Stephen Taylor**, New Jersey Institute of Technology, United States

Techniques are developed to estimate the present day value of the future social security benefits of a retiree based upon their chosen date of retirement, the term structure of interest rates, and life expectancy forecasts. These valuation methods are then used to determine the optimal retirement time of a beneficiary given a specific wage history and health profile in the sense of maximizing the present day value of future cashflows. We then examine how a number of risk factors including interest rates, disease diagnosis, and population life table risks impact the current value of

future payments. Specifically, we utilize principal component analysis in order to assess interest rate and population life expectancy variation risks. We then examine how such risks range over distinct income and demographic groups and finally summarize future research directions.

C1672: **Optimisation of trading strategies based on implied volatility and implied dividend volatility**
Presenter: **Enoch Nii Boi Quaye**, University of Kent, United Kingdom
Co-authors: Radu Tunaru

The aim is to compare information on implied volatility surface of stock-index to the corresponding implied volatility surface of index dividend futures on the stock index. We outline a computational procedure for aggregating implied volatility estimates based on the Black-Scholes, Black model and the model-free approach. Our findings illustrate how implied volatility term-structure of STOXX 50 with time-to-maturity exceeding 9-months moves enough to be justified by subsequent dividend fluctuations. Options with maturities between 1-9 months lead to implied volatilities that move too much to be justified by forward looking changes in dividends. The implied volatility term-structure of stock consistently exceeds that of index dividend futures thereby confirming Shiller's dividend puzzle under novel financial data and instruments. However, the magnitude of excess implied volatility declines with long-dated time-to-maturity, suggesting that discrepancies between the two are influenced by investment horizon. In addition, we design a set of trading strategies using implied volatility and the ratio of implied volatilities as a trading signal that prove to be successful for STOXX50 equity space.

C1442: **Asset pricing with endogenous state-dependent aggregate risk aversion**
Presenter: **Rachida Ouysse**, University of New South Wales, Australia

An economy is studied where the risk aversion is stochastic and beliefs-dependent. We formulate a consumption-based asset pricing model in which aggregate risk aversion is state-dependent (SDRA) in response not only to news about aggregate consumption as in the habit formation model but also to news about a wide range of key economic indicators. The representative consumer forms their beliefs toward risk from information available about a large number of variables that describe the economy. High-dimensionality of the available information is handled using a factor model. We use the generalized method of moments to estimate a nonlinear model that links information about the states of economic booms and busts to the aggregate risk aversion and to the Euler equations of the pricing model. We explore the pricing implications for a certain cross-section of stock returns. The empirical results support the hypothesis that aggregate risk aversion is counter-cyclical and varies with news about the business cycle. We observe volatility clustering of movements of risk aversion around recession periods. In addition to the price of consumption risk associated with consumption risk (as in the standard consumption capital asset pricing model), the induced time variation in risk aversion introduces risk preferences as a new component in the risk premium due to co-variation between aggregate risk aversion and asset returns.

CO450 Room G4 FINANCIAL ECONOMETRICS Chair: Wenying Yao

C0441: **Detecting signed spillovers in Asia**
Presenter: **Moses Kangogo**, University of Tasmania, Australia
Co-authors: Vladimir Volkov

The dynamics of the signed-spillover across financial markets is analyzed by using historical decomposition. By incorporating Markov switching-framework into VAR model, we investigate the dynamics of the signed-spillover during period of turbulent and period of tranquillity. Additionally, this approach would detect the source and direction of the spillover as well as identify the sign effects of the spillover. By detecting the dynamics of the signed-spillover in different regimes, this approach would out-perform the classical single-regime framework. We apply the methodology into high frequency RV data for the sample period 1999-2017. Empirical finding shows that the spillover are intense during period of turbulent and moderate during period of tranquillity.

C1778: **Asymmetric network connectedness of fears**
Presenter: **Mattia Bevilacqua**, London School of Economics, United Kingdom
Co-authors: Jozef Barunik, Radu Tunaru

The purpose is to study how shocks to the forward-looking expectations of future stock prices, extracted from call and put options, create asymmetric network connections. We introduce a new measure of network connectedness, called asymmetric fear connectedness, which captures the information related to "fear" on both sides of the options market, and that can be a useful forward-looking systemic risk monitoring tool. The decomposed connectedness measures provide timely predictive information for near-future macroeconomic and uncertainty indicators, and they contain additional valuable information not included in the aggregate network connectedness measure. The role of a positive/negative "fear" transmitter/receiver emerges clearly when we focus on idiosyncratic events for financial institutions. We identify banks that are predominantly positive/negative receivers of "fear", as well as banks that positively/negatively transmit "fear" in the financial system.

C0308: **Cointegration in high frequency data: Estimation and test**
Presenter: **Simon Clinet**, Keio University, Japan
Co-authors: Yoann Potiron

A framework is considered which adapts the notion of cointegration when two asset prices are generated by a driftless Ito-semimartingale featuring jumps with infinite activity, observed synchronously and regularly at high frequency. We develop a regression based estimation of the cointegrated relations method and show the related consistency and central limit theory when there is cointegration within that framework. We also provide a Dickey-Fuller type residual based-test for the null of no cointegration against the alternative of cointegration, along with its limit theory. Under no cointegration, the asymptotic limit is the same as that of the original Dickey-Fuller residual based test, so that critical values can be easily tabulated in the same way. Finite sample indicates adequate size and good power properties in a variety of realistic configurations, outperforming original Dickey-Fuller and Phillips-Perron type residual based tests, whose sizes are distorted by non ergodic time-varying variance and power is altered by price jumps. Two empirical examples consolidate the Monte-Carlo evidence that the adapted tests can be rejected while the original tests are not, and vice versa.

C1369: **Disentangling sources of high frequency market microstructure noise**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Simon Clinet

Employing tick-by-tick maximum likelihood estimation on several leading models from the financial economics literature, we find that the market microstructure noise is mostly explained by a linear model where the trade direction, i.e. whether the trade is buyer or seller initiated, is multiplied by the dynamic quoted bid-ask spread. Although reasonably stable intraday, this model manifests variability across days and stocks. Among different observable high frequency financial characteristics of the underlying stocks, this variability is best explained by the tick-to-spread ratio, implying that discreteness is the first residual source of noise. We determine the bid-ask bounce effect as the next source of noise.

C0230: **Estimating the rank of cojumps in high-dimensional financial data with market microstructure noise**
Presenter: **Wenying Yao**, Deakin University, Australia
Co-authors: Lars Winkelmann

Jumps explain a significant percentage of the overall variation of asset prices. The aim focuses on the reduced rank structure of contemporaneous jumps across assets. The cojump matrix is estimated from high-frequency financial data which are potentially corrupted by market microstructure noise. In general, noise can distort high-frequency statistics, such as volatility and jump estimates. We use the pre-average method to establish a

noise-robust estimator of the cojump matrix, which follows a mixed normal distribution. Then the estimation procedure of the cojump rank comes from Random Matrix Theory (RMT), employing the asymptotic results of spiked covariance matrix. This estimator is consistent even when the number of assets and the number of cojump events both diverge to infinity. We use this method to investigate the number of factors in the term structure of U.S. interest rates at macroeconomic news announcement times.

CO749 Room G5 FORECASTING AND ESTIMATION METHODS IN TIME SERIES ECONOMETRICS Chair: Sondre Holleland

C0369: **Measuring conditional dependence using the local Gaussian partial correlation**
Presenter: **Haakon Otneim**, Norwegian School of Economics, Norway
Co-authors: Dag Tjoestheim

It is well known that the dependence structure for jointly Gaussian variables can be fully captured using correlations, and that the conditional dependence structure in the same way can be described using partial and conditional correlations. The partial correlation does not, however, characterize conditional dependence in many non-Gaussian populations. We introduce the local Gaussian partial correlation (LGPC), a new measure of conditional dependence. It is a local version of the partial correlation coefficient that characterizes conditional dependence in a large class of populations. It has some useful and novel properties besides: The LGPC reduces to the ordinary partial correlation for jointly normal variables, and it distinguishes between positive and negative conditional dependence. Furthermore, the LGPC can be used to study departures from conditional independence in specific parts of the distribution. We provide several examples on this, both simulated and real, and derive estimation theory under a local likelihood framework. Finally, we indicate how the LGPC can be used to construct a powerful test for conditional independence, which, again, can be used to detect Granger causality in time series.

C0506: **Inflation: An MCMC estimator of the long-memory parameter in a state space model**
Presenter: **Fredrik NG Andersson**, Lund University, Sweden
Co-authors: Yushu Li

Inflation targeting is a common monetary policy regime. Inflation targets are often flexible in the sense that the central bank allows inflation to temporarily deviate from the target to avoid causing unnecessary volatility in the real economy. We propose modeling the degree of flexibility using an autoregressive fractionally integrated moving average (ARFIMA) model. Assuming that the central bank controls the long-run inflation rate, the fractional integration order becomes a measure of how flexible the inflation target is. A higher integration order implies that inflation deviates from the target for longer periods of time and consequently, that the target is flexible. Several estimators of the fractional integration order have been proposed in the literature. It has been shown that a state-based maximum likelihood estimator is superior to other estimators, but our simulations show that their finding is over-biased for a nearly non-stationary time series. To resolve this issue, we first proposed a Bayesian Monte Carlo Markov Chain (MCMC) estimator for fractional integration parameters. This estimator resolves the problem of over-bias. We estimate the fractional integration order for 6 countries for the period 1993M1 to 2017M9. We found that inflation was integrated to an order of 0.8 to 0.9 indicating that the inflation targets are implemented with a high degree of flexibility.

C0512: **The value of turning-point detection for optimal investment**
Presenter: **Michail Chronopoulos**, City, University of London, United Kingdom
Co-authors: Lars Sendstad

To capture the dynamic evolution of economic indicators and its impact on option pricing, we develop a regime-switching, real options framework for investment under uncertainty that facilitates time-varying transition probabilities. Considering a private firm with a perpetual option to invest, we use machine-learning techniques to forecast the evolution of transition probabilities and analyse how they affect the value of an investment opportunity. Results indicate that: (a) ignoring the dynamic evolution of transition probabilities can result in severe valuation errors; and (b) when the probability of a regime switch is low, the option value is greater in the good (bad) regime under time-varying than under fixed transition probabilities.

C1011: **Spatio-temporal ARMA-GARCH models**
Presenter: **Sondre Holleland**, University of Bergen, Norway
Co-authors: Hans Arnfinn Karlsen

The ARMA-GARCH time series model is expanded to a gridded spatio-temporal situation, with both spatial and temporal dependence. The spatial dependence is limited due to a user-specified translation invariant neighbourhood structure, for instance having only the closest neighbours influence a given point. This makes the model (potentially) stationary and sparse. The GARCH process contributes with conditionally heteroskedastic, uncorrelated innovations in the MA part of the ARMA model. Wrapping the neighbourhood system onto a torus is done to avoid a boundary problem in the conditional likelihood estimation. This finite, stationary model is called a circular one and is motivated from a projection of the unlimited version. We will introduce the models, dwell on some issues, discuss estimation and show some examples of when one should include a GARCH term in the spatio-temporal ARMA and when not to. A real data example of processed cell imagery data will be presented.

C1129: **Model specification testing in seasonal ARMA models**
Presenter: **Johan Lyhagen**, Uppsala University, Sweden

The joint asymptotic distribution of the coefficients of the autocorrelation function and the partial autocorrelation function is derived. Based on this, a test is proposed to test the specification of a seasonal ARMA model. The test is shown, by simulations, to have superior properties compared to testing the residuals being white noise.

CO312 Room Montague PRICING KERNELS AND FACTOR MODELS Chair: Benjamin Holcblat

C0262: **Correcting misspecification in stochastic discount factor models**
Presenter: **Irina Zviadadze**, HEC Paris, France
Co-authors: Raman Uppal, Paolo Zaffaroni

It is shown how, given a misspecified stochastic discount factor (SDF), one can construct an admissible SDF, namely an SDF that prices assets correctly. We first extend the traditional Arbitrage Pricing Theory (APT) to capture misspecification from both pervasive (systematic) pricing errors and idiosyncratic pricing errors. The constructed admissible SDF, which uses the extended APT as its foundation, satisfies the already-known bound exactly. If the number of assets N is large, the admissible SDF recovers the contribution of the missing pervasive factors completely without requiring one to identify the missing factors. Indeed, projecting the correction term of the SDF on the space spanned by the candidate missing factors, achieves an R^2 that converges to one as N increases. Simulations demonstrate that the theory we develop is remarkably effective in correcting various sources of misspecification.

C0263: **An information-theoretic asset pricing model**
Presenter: **Christian Julliard**, London School of Economics, United Kingdom

A non-parametric estimate of the pricing kernel, extracted using an information-theoretic approach, is shown to deliver smaller out-of-sample pricing errors and a better cross-sectional fit than leading factor models. The information SDF (I-SDF) identifies sources of risk not captured by standard factors, generating very large annual alphas (10%-18%) and Sharpe ratios (0.90-1.3). I-SDFs extracted from a wide cross-section of equity

portfolios are highly positively skewed and leptokurtic, and imply that about half of the observed risk premia represent a compensation for tail risk. The I-SDF offers a powerful benchmark relative to which competing theories and investment strategies can be evaluated.

C0425: **Empirical evaluation of overspecified asset pricing models**
Presenter: **Enrique Sentana**, CEMFI, Spain
Co-authors: Francisco Penaranda, Elena Manresa

Asset pricing models with potentially too many risk factors are increasingly common in empirical work. Unfortunately, they can yield misleading statistical inferences. Unlike other studies focusing on the properties of standard estimators and tests, we estimate the sets of SDFs and risk prices compatible with the asset pricing restrictions of a given model. We also propose tests to detect problematic situations with economically meaningless SDFs uncorrelated to the test assets. We confirm the empirical relevance of our proposed estimators and tests with the linearized version of the consumption CAPM, and provide Monte Carlo evidence on their reliability in finite samples.

C0355: **Anomaly or risk factor? Some simple tests**
Presenter: **Abraham Lioui**, EDHEC Business School, France
Co-authors: Michael Weber, Benjamin Holcblat

Hundreds of factors predicting cross-sectional returns have been discovered. We develop simple tests to assess whether risk can explain the predicting power of these factors. Our tests account for all kinds of risk disliked by risk-averse individuals, including high-order moments and tail risk. Our tests do not rely on the validity of a factor model nor other parametric statistical model.

C0347: **Deep learning in asset pricing**
Presenter: **Markus Pelger**, Stanford University, United States
Co-authors: Jason Zhu, Luyang Chen

We propose a novel approach to estimate asset pricing models for individual stock returns that takes advantage of the vast amount of conditioning information, while keeping a fully flexible form and accounting for time-variation. Our general non-linear asset pricing model is estimated with deep neural networks applied to all U.S. equity data combined with a substantial set of macroeconomic and firm-specific information. We estimate the stochastic discount factor that explains all asset returns from the conditional moment constraints implied by no-arbitrage. Our asset pricing model outperforms out-of-sample all other benchmark approaches in terms of Sharpe ratio, explained variation and pricing errors. We trace its superior performance to including the no-arbitrage constraint in the estimation and to accounting for macroeconomic conditions and non-linear interactions between firm-specific characteristics. Our generative adversarial network enforces no-arbitrage by identifying the portfolio strategies with the most pricing information. Our recurrent Long-Short-Term-Memory network finds a small set of hid- den economic state processes. A feedforward network captures the non-linear effects of the conditioning variables. Our model allows us to identify the key factors that drive asset prices and generate profitable investment strategies.

CO234 Room Woburn ASSESSING MACROECONOMIC POLICIES Chair: Nora Traum

C0199: **Government spending multiplier and the size of the shock: Evidence from U.S.**
Presenter: **Madina Karamysheva**, Higher School of Economics, Russia
Co-authors: Nikita German

The aim is to investigate whether fiscal multiplier depends negatively on the size of the government spending shock. We build our hypothesis on behavioral arguments and check it empirically on US data. For doing so, we adopt state-dependent VAR, accompanied by Jorda local projections method, and show that investigated relationship is U-shaped: for small shocks in government consumption and investment, the fiscal multiplier is rising in size of the shock, while for large ones it falls. We address possible endogeneity issues and illustrate that our results are non-sensible to these concerns. Finally, we limit our analysis to government consumption multiplier, as our hypothesis suggests strong non-constancy namely in this position. We find a strong negative relationship between government consumption multiplier and the size of the shock.

C0206: **Labor adjustment and productivity in the OECD**
Presenter: **Vivien Lewis**, Deutsche Bundesbank, Germany
Co-authors: Maarten Dossche, Andrea Giovanni Gazzani

Labor productivity is more procyclical in countries with lower unemployment volatility. To capture this new stylized fact, we propose a business cycle model with firing costs, variable hours and effort. Effort helps to capture the procyclicality of labor productivity in the presence of demand shocks. A structural reform that reduces firing costs leads to a greater expansion of employment when effort can vary. Moreover, labor market deregulation makes labor productivity less procyclical.

C0543: **Origins of macro policy shifts: A new approach to regime switching in DSGE models**
Presenter: **Boreum Kwak**, Martin Luther University Halle-Wittenberg and Halle Institute for Economic Research, Germany
Co-authors: Yoosoon Chang, Bing Li, Fei Tan

The origins of aggregate fluctuations in the U.S. data are investigated by using Bayesian analysis of a New-Keynesian DSGE model with monetary-fiscal policy interactions. We introduce regime switching into the model that links the current regime of the economy to the past structural shocks by an autoregressive regime factor. Using a simple analytical model, such linkage is shown to generate endogenous feedback from the behavior of underlying economic fundamentals to the regime generating process. Our key empirical findings are twofold. First, non-policy shocks, most notably the markup shock, have played a predominant role in driving regime changes during the post-World War II era. Second, the Bayes factor strongly favors the endogenous switching version of the model over the exogenous case. We conclude that endogenizing regime changes in DSGE models with monetary-fiscal policy interaction provides both a theoretically and empirically promising venue for understanding the purposeful nature of policy interaction.

C1886: **Fiscal multipliers with an informal sector**
Presenter: **Evangelia Vourvachaki**, Bank of Greece, Greece
Co-authors: Dimitris Papageorgiou, Dimitris Malliaropulos, Harris Dellas

The shadow economy exaggerates the effect of fiscal policy on both official and true economic activity. We measure these effects using a model with an informal sector in the context of the recent 2010-2015 Greek fiscal consolidation experience. We find that formal output declined by 50% more than projected (26% vs 18%) whereas, due to the large increase in the share of the informal sector (by 50%), true output declined by much less (17%). Almost one third of the formal GDP decline was due to income tax rate increases (which failed to raise extra tax revenue). Our model predicts that had the informal sector been contained to its pre-crisis size, at least one-quarter of the decline in GDP could have been averted. And that the capital controls imposed in 2015 may inadvertently have, within a year, contributed to a reduction of the share of the informal sector by almost five percentage points and to an increase in formal GDP by 2.6%.

C1887: **Fiscal distress and banking performance: The role of macroprudential regulation**
Presenter: **Dimitris Papageorgiou**, Bank of Greece, Greece
Co-authors: Hiona Balfoussia, Harris Dellas

Fiscal fragility can undermine a governments ability to honor its bank deposit insurance pledge and induces a positive correlation between sovereign

default risk and financial (bank) default risk. We show that this positive relation is reversed if bank capital requirements in fiscally weak countries are allowed to adjust optimally. The resulting higher requirements buttress the banking system and support higher output and welfare relative to the case where macroprudential policy does not vary with the degree of fiscal stress. Fiscal tenuousness also exacerbates the effects of other risk shocks. Nonetheless, the economy's response can be mitigated if macroprudential policy is adjusted optimally. Our analysis implies that, on the basis of fiscal strength, fiscally weak countries would favor and fiscally strong countries would object to banking union.

CO404 Room Jessel TIME SERIES ECONOMETRICS: NONSTATIONARITIES AND INSTABILITIES Chair: Martin Wagner

C1446: **Monitoring structural breaks of the cointegration rank in error correction models**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

Consistent monitoring procedures are developed with the goal to detect structural changes in a Johansen type error correction model. In particular, we consider breaks where the cointegration rank remains constant as well as breaks changing the cointegration rank. Lagrange multiplier tests are developed allowing to monitor these kinds of breaks. The monitoring procedure is used to investigate possible structural changes the risk-adjusted forward unbiasedness hypothesis.

C1542: **Pseudo maximum likelihood analysis of I(2) processes in the state space framework**
Presenter: **Lukas Matuschek**, Technical University Dortmund, Germany
Co-authors: Dietmar Bauer, Patrick de Matos Ribeiro, Martin Wagner

Nominal macroeconomic time series are regularly found to be adequately described as $I(2)$ processes, with cointegration analysis typically performed in the vector autoregressive (VAR) framework. The VAR framework may be too restrictive: First, VAR processes are not closed under marginalization or aggregation, where in both cases the resulting processes are in general vector autoregressive moving average (VARMA) processes. Second, the solutions of dynamic stochastic economic models are typically VARMA processes rather than VAR processes. To overcome the limitation to VAR processes we develop estimation and inference techniques for $I(2)$ cointegrated VARMA processes cast in state space format. In particular we derive consistency as well as the asymptotic distributions of estimators maximizing the Gaussian pseudo likelihood function. As usual, the parameters corresponding to $I(2)$ and $I(1)$ variables are estimated super-consistently at rates T^2 and T respectively, whereas all other parameters are estimated at rate $T^{1/2}$. The limiting distributions of the parameters corresponding to the integrated components are mixtures of Brownian motions, the parameters of the stationary subsystem are asymptotically normally distributed. Furthermore, we discuss hypothesis tests for the cointegrating ranks as well as for the cointegrating spaces.

C1543: **Stochastic trends and economic fluctuations reconsidered**
Presenter: **Patrick de Matos Ribeiro**, Technical University Dortmund, Germany
Co-authors: Dietmar Bauer, Lukas Matuschek, Martin Wagner

In a seminal paper the empirical performance of the neoclassical stochastic growth model by vector autoregressive (VAR) cointegration analysis has been investigated by using quarterly observations from 1949:1 to 1988:4. In addition to the real variables, private consumption, private investment and GNP net of government spending, they also consider nominal quantities, i.e., money (M2), prices (GNP deflator) and a short-term interest rate (federal funds rate). Economic theory predicts three cointegrating relationships, the two great ratios (consumption-ouptut and investment-output) as well as a money demand relationship. Furthermore, the relative importance of permanent and transitory shocks for the dynamic behaviour of the US economy is investigated. As is well-known, the solutions of dynamic stochastic economic models are generally vector autoregressive moving average rather than VAR processes, with the solutions typically given in state space format. Building upon recent advances in state space cointegration analysis, this paper reassesses, and extends by using data potentially up to 2018:4, the KPSW analysis from a state space perspective.

C1810: **Right matrix fraction description for stochastically singular models: Structure theory and estimation**
Presenter: **Juho Koistinen**, University of Helsinki, Finland
Co-authors: Bernd Funovits

The focus is on the estimation of stochastic processes with rational spectral densities that are rank deficient. Interest in these processes has emerged in relation to generalized dynamic factor models (GDFMs), which contain fewer economic shocks than endogenous variables. Our contribution is to propose a right matrix fraction description (MFD) realization of the transfer function, $\chi_t = k(z)\varepsilon_t = d(z)c(z)^{-1}\varepsilon_t$ with $d(z) \in \mathbb{R}^{N\times q}$ and $c(z) \in \mathbb{R}^{q\times q}$ which has two advantages over the usual two-step estimation procedure (where firstly a static transformation reduces to a static factor process with full rank covariance matrix and secondly a VAR model with fewer inputs than outputs is estimated): Its one-step nature makes it potentially more efficient and the same dynamics might be modeled with fewer parameters. The right MFD realization allows for more generality compared to the singular VAR since the covariance of the N-dimensional common factors χ_t at lag 0, $\mathbb{E}\chi_t\chi_t'$, can be of rank $r < N$. Moreover, the rank of the spectral density of the common factors is $q < r$. We analyse the properties of the column Kronecker canonical form (which can be written as a minimal state space realization), and highlight its usefulness for modeling the common factor of GDFMs. Using the state space representation of the right MFD, estimation can be performed efficiently via Kalman filtering.

C1843: **Flexible nonlinear trend specifications and cointegration**
Presenter: **Hanno Reuvers**, Erasmus University Rotterdam, Netherlands
Co-authors: Yicong Lin

Inference is developed for a model that combines a power law trend specification with a polynomial cointegration framework. We provide the asymptotic distribution of the nonlinear least squares (NLS) estimator when the regression errors and integrated regressors are serially dependent and cross correlated. The limiting distribution is nonstandard thereby complicating inference. We discuss two methods to conduct inference: (1) fully modified estimation, and (2) simulated estimation of the null distribution. An extensive simulation study illustrates and compares the finite sample performance of these methods. Our approach is illustrated with a detailed study on the existence of the Environmental Kuznets Curve (EKC) for Belgium, Denmark, France, Netherlands, UK and USA over the period 1870-2014. The main question at hand is whether stochastic trends or deterministic trends are responsible for the nonlinearities observed in the data.

CC827 Room MAL 351 CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: Marc Gronwald

C1749: **Value added in Italian primary school: An econometric approach**
Presenter: **Marinella Boccia**, University of Salerno, Italy
Co-authors: Adalgiso Amendola, Alessandra Amendola

To find models that provide a valid measure of the effectiveness of the school quality and of education value added is crucial for the accountability. The aim is to compute a value added measure comparing different econometric models and investigate on which of them performs better. In order to do that the study will use the INVALSI standardized tests in Italian and math for primary school. The preliminary results make in evidence that higher test scores are driven by the test scores in the previous period, such as by the kindergarten attendance. The parents educational and professional level also reveals its significance. Remarkable, negative and significant is the coefficient indicating the presence of monitoring. However the most important tool derived is the school value added distribution, plotted in the graphs, realized both for area and representative

regions. According to these the south and the regions there located show a lower value of value added and a higher density on the left and side of distribution. Differently from that the north and some of its regions exhibit a more symmetric value added distribution with higher values.

C1799: **A wavelet analysis of the ripple effect in UK regional house prices**
Presenter: **Iolanda Lo Cascio**, University of Palermo, Italy

The aim is to gain insights on the spatio-temporal mechanism of house price spillovers, also known as ripple effect, among 13 UK regional housing markets, over the period 1973 2018. From a policy perspective, it might be essential to discriminate if the effects of a shock decay more slowly along the geographical dimension as compared to the decay along the time dimension. We contribute to the debate by focusing also on whether the ripple effect varies across different phases of the housing cycles, i.e., if there is a strong comovement when the housing market is prosperous or otherwise. We enter the debate in a novel manner, by using some wavelet analysis tools (wavelet coherence and phase differences amongst others), which reveal the spectral characteristics of a series and show how different periodic components of the prices evolve over time. Results are interesting. Price spillovers from London to other markets are detected both in the short and in the long run for those regions closer to London. More distant regions price changes lead London market only in the short run and for selected time spans; however, in the long run London remains the dominant market.

C1670: **Relationship between money supply and other macroeconomic variables in the Philippines: A vector autoregression analysis**
Presenter: **Rutcher Lacaza**, University of the Philippines, Philippines
Co-authors: Barton Sy

Using quarterly data, a vector autoregressive (VAR) model was used to examine the relationship between money supply (M2) and other macroeconomic variables in the Philippines particularly, real GDP, inflation and interest rate starting at the period when the Bangko Sentral ng Pilipinas (BSP) formally adopted the inflation targeting framework in January 2002. Granger causality test, impulse response functions and variance decomposition were also implemented to analyze and uncover the impact of these macroeconomics variables on the money supply. After performing an empirical analysis, the result shows that a rise in the real GDP can cause an increase in the money supply; similarly, a rise in the inflation rate can result to an increase in the money supply; conversely, a rise in the interest rate can lead to a decrease in the money supply. Shocks on the significant macroeconomic variables have a significant effect on money supply at certain periods. The variance decomposition of all models shows that a significant proportion of the movements in money supply is due to shocks on the other variables especially at short horizons. The results suggest that policy makers in the Philippines can have a better control of money supply by adjusting the real GDP, inflation rate and interest rate.

C1844: **Non-parametric testing of information asymmetry in the U.S. mortgage servicing market**
Presenter: **Helmi Jedidi**, HEC Montreal, Canada
Co-authors: Georges Dionne

The main objective is to test for evidence of information asymmetry in the U.S. mortgage servicing market. The main research question is: does selling the mortgage servicing rights by the initial lender to a second servicing institution unveil any residual asymmetric information? We investigate the link between the originators decision to sell the mortgage underlying MSR and the mortgage default likelihood using a large sample of U.S. mortgages that were securitized through the private-label channel during the 2000-2013 period. Our econometric methodology is purely non-parametric. We use Kernel Density Estimation technique to estimate the multivariate conditional density function. Our findings support the presence of a second-stage asymmetric information in the U.S. mortgage servicing market. Our empirical results show a significant positive relationship between the lenders decision to sell the underlying MBS and mortgage default. Our evidence suggests that originating lenders are indeed taking advantage of privileged information they obtain at the time of original underwriting. We also use parametric tests to corroborate our results after controlling for observable risk characteristics, econometric misspecification error, and endogeneity issues using instrumental variables approaches and simultaneous equations.

C0729: **Modeling the UK mortgage demand using online searches**
Presenter: **Jaroslav Pavlicek**, Institute of Economic Studies, Charles University, Czech Republic
Co-authors: Ladislav Kristoufek

Internet has become the primary source of information for most of the population in modern economies and as such it provides enormous amount of readily available data and the data on the internet search queries have been shown to improve forecasting models for various economic and financial series. In the aftermath of the global financial crisis, modeling and forecasting mortgage demand has become a central issue in the banking sector as well as for governments and regulators. In the UK, the mortgage market dynamics is could be measured by new mortgage approvals. As the online searches are expected to be one of the last steps before the actual customer application for a large share of population, the intuitive utility of utilizing the intensity of specific online search queries to model them is appealing. When comparing two baseline models - an autoregressive model and a structural model with relevant macroeconomic variables - with their extensions utilizing online searches on Google, the extended models show to better explain the number of new mortgage approvals and improve their nowcasting and forecasting performances markedly. Moreover, utilizing machine learning techniques, the data on Google searches are preferred and, to a certain extent, able to replace the macroeconomic indicators.

CG225 Room G21A CONTRIBUTIONS IN MACROECONOMICS AND MACROECONOMETRICS Chair: Luis Aguiar-Conraria

C0481: **Endogenous time-variation in vector autoregressions**
Presenter: **Danilo Leiva-Leon**, Banco de España, Spain

An econometric framework is proposed that provides robust inference on the origins of instabilities in the relationship between key macroeconomic variables. We introduce a new class of Time-Varying Parameter Vector Autoregression (TVP-VAR) models where the set of underlying structural shocks are allowed to potentially influence the dynamics of the autoregressive coefficients. The proposed Endogenous TVP-VAR framework is applied to study the sources of instabilities in the relationship between the unemployment, inflation and interest rates of the U.S. economy. The results indicate that cost-push shocks are an important source of macroeconomic instability and emphasize the role of lags in the transmission mechanism of monetary policy.

C1589: **On the estimation of behavioral macroeconomic models via simulated maximum likelihood**
Presenter: **Jiri Kukacka**, Charles University, Faculty of Social Sciences, Czech Republic
Co-authors: Tae-Seok Jang, Stephen Sacht

The simulated maximum likelihood estimation method is extended to multivariate macroeconomic optimization problems and employ it to identify the behavioral heuristics of heterogeneous agents in the baseline three-equation New Keynesian model. This approach considerably relaxes restrictive theoretical assumptions and enables a novel estimation of the intensity of choice parameter in the discrete choice switching process. Using Monte Carlo simulation, we first analyze the properties of the estimation framework and study its ability to consistently recover the pseudo-true parameters in a controlled environment. The proposed method favors estimation of the switching parameter; however, the curse of dimensionality arises via a consistent downward bias for idiosyncratic shocks. Our empirical results show that the forward-looking version of both the behavioral and the rational model specifications exhibits good performance. We further identify potential sources of misspecification for the hybrid version. A novel feature of our analysis is that we pin down the switching parameter for the intensity of choice for the Euro Area and US economy.

C1855: **A Bayesian ECM for forecasting real ruble exchange rate under structural change in the monetary policy**
Presenter: **Nikita Fokin**, Russian Presidential Academy of National Economy and Public Administration, Russia

Co-authors: Andrey Polbin

At the end of 2014, the Central Bank of Russia changed the monetary policy regime. The goal is to construct the model, which can accurately forecast the real exchange rat immediately after changing the monetary policy regime. Following the assumption of long-term neutrality of money, we assume that the transition to a new regime has changed the mechanism of adapting of the real exchange rate to long-term equilibrium in response to the shock of oil prices, but the long-term parameters of the model have not changed a lot. First, we estimate the frequentist ECM on the period of the old monetary policy regime. The estimates of parameters of the cointegration relation and their confidence intervals are set as priors for long-term parameters of the equation of the new regime. For short-term parameters, uninformative priors are set. Bayesian specification of the model, in contrast to the frequentist model, has two advantages. The first one is that it allows us to predict the real exchange rate immediately after the monetary policy regime was changed, when for the frequentist model it is necessary to collect a minimum set of observations to estimate the parameters of the new regime. The second advantage is that we allow long-term parameters to change slightly due to a change in the monetary policy regime, when in the frequentist model we would have to impose a strict restriction on the equality of the parameters of the cointegration relation in different regimes.

C0275: **Study of sovereign credit rating determinants: A Bayesian averaging model**

Presenter: **Zied Ftiti**, EDC Paris Business School, France

The impact of the debt crisis on sovereign credit rating assessment in 24 European countries is examined over the period spans from 2005 to 2016. Due to the heterogeneity that exists between countries of the sample, we separate the sample into two sub-samples: GIIPS and off-GIIPS. Given the uncertainty issue related to estimated parameters as well as model space, we set up a Bayesian Averaging Model. The results suggest that unemployment, inflation and DEBT per GDP ratio are common and robust factors behind rating during both periods and for two groups of countries. We find also that external debt, governance indicators and GDP per capita seem to be significant during both periods in GIIPS, however, their importance becomes significant for the second group during post-crisis only. The results support the criticisms delivered to rating agencies concerning their methodology's instability, pro-cyclical behavior that leads to their failure to anticipate a crisis.

C1974: **The role of economic and financial indicators in forecasting stock volatility**

Presenter: **Aya Ghalayini**, Lancaster University, United Kingdom

Co-authors: Marwan Izzeldin

Financial volatility is an essential input for investors and policymakers alike. Therefore, the different macroeconomic and financial determinants that drive financial volatility are emphasized. We revisit the relationships between stock market volatility and different factors using recent models that allow for the combination of variables with different frequencies. It is shown that HARX using Chow-Lin interpolations method supersedes MFVAR in terms of forecastability of the financial volatility. The study is performed on a multi-dimensional scale where three forecasting horizons (daily, weekly, and monthly), three regimes (pre-crisis, crisis, and post-crisis), two volatility measures (RV and BV), and three loss functions (MSE, HMSE, and MAE) are examined. The data involves SPY and two stocks from each of the ten business sectors of the US economy.

CG025 Room MAL 352 CONTRIBUTIONS IN ECONOMETRICS OF VOLATILITY Chair: Roman Liesenfeld

C1717: **Singular conditional autoregressive Wishart model**

Presenter: **Gustav Alfelt**, Stockholm University, Sweden

Co-authors: Taras Bodnar, Farrukh Javed, Joanna Tyrcha

A Singular Conditional Autoregressive Wishart model that aims to capture the dynamics of singular realized covariance matrices of asset returns is suggested. Such singularity arises in high-dimensional cases where the dimension of the return process exceeds the number of intraday returns sampled each day. The model assumes that the non-singular scale matrix of the underlying Singular Wishart process follows an autoregressive moving average structure with a BEKK specification, and can be estimated by the Maximum Likelihood method. In order to facilitate feasible estimation in high-dimensional cases, the model is fitted to a transformation of the data series together with the application of covariance targeting. Finally the model is applied to high-frequency data from AMEX, NASDAQ and NYSE, and is evaluated by out-of-sample forecast accuracy.

C1832: **Realized volatility estimator under liquidity constraints**

Presenter: **Erindi Allaj**, Epoka University, Albania

The behaviour of the realized volatility (RV) estimator is analyzed under liquidity constraints. The liquidity is measured by the impact of the trade size on the asset price. We find that this estimator is inconsistent but convergent in probability. Motivated by this fact, we propose a new estimator which is consistent and asymptotically unbiased under liquidity constraints. Finally, our results are validated by a simulation study.

C1875: **Parameters identification for inverse option problems using Markov Chain Monte Carlo methods**

Presenter: **Yasushi Ota**, Okayama University of Science, Japan

The inverse option problems (IOP) in the extended Black–Scholes model arising in financial market are investigated. We identify the volatility and the drift coefficient from the measured data in financial markets using a Bayesian inference approach, which is presented as an IOP solution. The posterior probability density function of the parameters is computed from the measured data. The statistics of the unknown parameters are estimated by a Markov Chain Monte Carlo (MCMC) algorithm, which exploits the posterior state space. The efficient sampling strategy of the MCMC algorithm enables us to solve inverse problems by the Bayesian inference technique. Our numerical results indicate that the Bayesian inference approach can simultaneously estimate the unknown trend and volatility coefficients from the measured data.

C1898: **An asymmetrical opposite-signed-shocks GARCH model**

Presenter: **Andrei Kostyrka**, University of Luxembourg, Luxembourg

Co-authors: Dmitry Malakhov

A new general GARCH-like framework is proposed. Asset returns are decomposed into a sum of copula-connected unobserved positive and negative shocks, possibly with discrete jumps (yielding up to 4 distinct shocks, continuous and discrete, of both signs). We model return distributions in a flexible manner using several parametric families of signed shocks and copulae with dynamic parameters. The model subsumes the Bad environments, good environments model as a special case. We compare our models with 40 well-established GARCH variants by estimating and backtesting them on a total of 19.5 years of S&P500 daily data. Our models perform better both out of sample (according to Christoffersen and Engle-Manganelli VaR forecast tests) and in sample (according to the non-nested LR tests and information criteria). For Diebold-Mariano forecast accuracy tests for volatilities, the picture is mixed; however, a subset of our models that have the best in-sample results and pass the VaR tests produces forecasts that are not worse than those of the GARCH variants that pass the same tests. Using these models, we reveal the information structure of returns and investors behaviour, e.g., market reaction to positive and negative shocks. Given the complexity of the new models, we use multiple numerical stabilisation techniques, including fail-safe numerical integration, stochastic and deterministic optimisation, robustified numerical derivatives, and parallel-capable estimation routines.

C1428: **A semi-parametric realized joint value-at-risk and expected shortfall regression framework**

Presenter: **Chao Wang**, The University of Sydney, Australia

Co-authors: Richard Gerlach

A new realized joint Value-at-Risk (VaR) and expected shortfall (ES) regression framework is proposed, through incorporating a measurement equation into the original joint VaR and ES regression model. The measurement equation models the contemporaneous dependence between the

realized measure (e.g. Realized Variance and Realized Range) and the latent conditional quantile. Further, sub-sampling and scaling methods are applied to both the realized range and realized variance, to help deal with inherent micro-structure noise and inefficiency. An adaptive Bayesian Markov Chain Monte Carlo method is employed for estimation and forecasting, whose properties are assessed and compared with maximum likelihood estimator through simulation study. In a forecasting study, the proposed models are applied to 7 market indices and 2 individual assets, compared to a range of parametric, non-parametric and semi-parametric models, including GARCH, Realized-GARCH, conditional autoregressive Expectile, and joint VaR and ES quantile regression models, one-day-ahead Value-at-Risk and Expected Shortfall forecasting results favor the proposed models, especially when incorporating the sub-sampled Realized Variance and the sub-sampled Realized Range in the model.

CG848 Room MAL 353 CONTRIBUTIONS IN BUSINESS CYCLE ANALYSIS Chair: Christopher Otrok

C0753: **Sales and promotions and the great recession deflation**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus
Co-authors: Christos Savva, Nektarios Michail, Niki Papadopoulou

The effect of sales and promotions on the pricing decisions of firms is investigated. A theoretical model is provided where firms face menu costs when adjusting their price and apply sales offers that decrease temporarily the listed price to attract higher demand, especially because households exert effort to locate the price deals. Thus, each period the final price is determined by the price set by the firm which is common knowledge to all agents and a sales deal that is a draw from a distribution with endogenous time-varying support. In a recession, even though prices in the economy look sticky, firms increase the frequency and the range of sales on their products substantially. This implies that traditional inflation measures are overstated in recessions, because they ignore the surge in sales and promotions and the consumers' tendency to hunt those limited time offers more actively. This framework can explain the mild deflation experienced during the Great Recession. Moreover, it is demonstrated that using traditional inflation measures can prolong recessions.

C1650: **Score-driven time-varying transition probabilities in a dynamic factor Markov-switching model**
Presenter: **Bram van Os**, Econometric Institute, Erasmus University Rotterdam, Netherlands
Co-authors: Dick van Dijk

The business cycle is an important driver of many macroeconomic variables. The existing dynamic factor Markov-switching (DFMS) model has proven to be a powerful framework to measure the cycle. This model estimates the latent business cycle factor by exploiting the cross-sectional information in multiple observed variables. Furthermore, in line with the macroeconomic intuition of expansions and contractions phases in the cycle the evolution of the factor is allowed to be regime-dependent, with a hidden Markov process dictating the regime-switches. Allowing for time-varying parameters in univariate Markov-switching models has been found useful in the context of business cycle applications. In particular, previous literature has amounted substantial evidence that the assumption of time-invariant transition probabilities used may not be appropriate here. The authors enhance the DFMS model to allow for time-varying transition probabilities (TVTP) by combining the accelerated score-driven framework with a method for adding score-driven TVTP to a Markov-switching model, which they extend to the multivariate setting and where they allow for exogenous variables. In an empirical application using the four components of The Conference Boards Coincident Economic Index for the period 1959-2019, it is found that the proposed framework allows for superior dating of US business cycle peaks and troughs.

C1814: **Modeling rent seeking activities: Quality of institutions, macroeconomic performance and the economic crisis**
Presenter: **Tryfonas Christou**, Athens University of Economics and Business, Greece
Co-authors: Vanghelis Vassilatos, Apostolis Apostolis Philippopoulos

The implications of institutional quality on macroeconomic performance is studied by augmenting the standard real business cycle model with rent-seeking competition. The idea is that agents allocate a part of their effort time competing with each other for a fraction of a contestable prize. We consider alternative contestable prizes like government transfers, tax revenues and firms' output, evaluate their ability to match the data and compare second-moment properties in the data vis-a-vis each model. Motivated by changes in government policy instruments after the 2007-8 world crisis, we investigate how this affected macroeconomic performance and institutional quality. Main findings: first, the repercussions of the crisis have been milder in countries with better institutional quality and second, countries with poor institutional quality pre-crisis, suffered further deterioration. All models perform in a similar way; however qualitative and quantitative differences arise in second-moment properties among core and periphery countries.

C0939: **The Phillips Curve at 60: Time for time and frequency**
Presenter: **Maria Joana Soares**, University of Minho, Portugal
Co-authors: Luis Aguiar-Conraria, Manuel Martins

The U.S. New Keynesian Phillips Curve is estimated in the time-frequency domain with continuous wavelet tools, to provide an integrated answer to the three most controversial issues on the Phillips Curve. (1) Has the short-run tradeoff been stable? (2) What has been the role of expectations? (3) Is there a long-run tradeoff? First, we find that the short-run tradeoff is limited to some specific episodes and short cycles and that there is no evidence of nonlinearities or structural breaks. Second, households' expectations captured trend inflation and were anchored until the Great Recession, but not since 2008. Then, inflation over-reacted to expectations at short cycles. Finally, there is no significant long-run tradeoff. In the long-run, inflation is explained by expectations.

C1669: **International information flows, sentiments and cross-country business cycle fluctuations**
Presenter: **Jacek Kotlowski**, SGH Warsaw School of Economics, Poland
Co-authors: Michal Brzoza-Brzezina, Grzegorz Wesolowski

Business cycles are strongly correlated between countries. One possible explanation (beyond traditional economic linkages like trade or finance) is that consumer or business sentiments spread over boarders and affect cyclical fluctuations in various countries. We first lend empirical support to this concept by showing that sentiments travel between countries at a speed much higher than can be explained by traditional linkages. Then we construct a two-economy new Keynesian model where noisy international information can generate cyclical fluctuations (co-movement of GDP, consumption, investment and inflation) in both countries. Estimation with US and Canadian data reveals a significant role of international noise shocks in generating such common fluctuations - for instance they explain between 15-30% of consumption variance in the US and Canada. Finally we show that our estimated noise shocks can indeed be interpreted as sentiment shock.

CG217 Room MAL 354 CONTRIBUTIONS IN MACHINE LEARNING IN FINANCE Chair: Michael Ellington

C1432: **Evolutionary relaxed support vector regression for exchange rates trading**
Presenter: **Shaolong Sun**, Xi'an Jiaotong University, China
Co-authors: Shouyang Wang, Yunjie Wei

A new evolutionary learning approach, namely CS-RSVR, is proposed for exchange rate forecasting and trading. The proposed CS-RSVR approach can dynamically optimize the values of all SVR's parameters through the CS evolutionary algorithm, and use acquired parameters to construct optimized RSVR in order for proceeded forecasting foreign exchange rates. Many researchers have discussed exchange rate forecasting with the majority focusing on forecasting performance, however; accuracy is only one part of exchange rate forecasting. More important is how integrated approaches such as this can guide professional practice. We extend our forecasting to test trading performance of exchange rates between the USD and four other major currencies, EUR, GBP, CNY and JPY. The experimental results demonstrate the CS-based optimized SVRs models

significantly improve efficient in trading terms compared with other optimized SVRs models. Generally speaking, our proposed CS-RSVRb model can be considered as a promising solution for exchange rates forecasting and trading.

C1507: **Industry return predictability: Evidence from China**

Presenter: **Yawen Zheng**, University of Liverpool, United Kingdom

Co-authors: Michael Ellington, Michalis Stamatogiannis

The purpose is to demonstrate that Chinese industry portfolio returns are able to forecast the returns of other industry portfolios. The forecastability differs between industry portfolios and some industries are more useful predictors than others. We use a machine learning technique (LASSO) to select the most relevant predictors for each industry and then analyse these relationships at monthly, weekly and daily data frequencies. We find significant cross-industry return forecastability and out-of-sample tests indicate positive forecasting performance. Our results hold to a number of robustness checks.

C1764: **Loan default analysis in Europe: Tracking regional variations using big data**

Presenter: **Luca Barbaglia**, European Commission Joint Research Centre, Italy

Co-authors: Sebastiano Manzan, Elisa Tosetti

The loan default behaviour in the European market is empirically investigated using a novel, big data set on over 20 million residential mortgages observed over the period from 2013 to 2018. We model the occurrence of a default as a function of loan-level information at origination, characteristics of the financial institution originating the loan, borrower's economic situation, as well as local economic conditions. We adopt three alternative machine learning techniques useful for predicting default events, namely the Logistic Regression, the Gradient Boosting and the XGBoost approaches, and carry the analysis at NUTS2 regional-level. We find that the most important variables in explaining default is the loan originator, the interest rate currently applied to the mortgage, and local economic characteristics, while other loan- or borrower-specific features are less relevant. We exploit techniques from a recent literature on interpretable machine learning to identify the most relevant factors affecting default and to capture the non-linear effects on default of some variables, like interest rate or changes in unemployment rate. Our results point at consistent geographical heterogeneity in variable importance magnitudes, indicating the need of European policy that is regionally tailored.

C1820: **A coherent framework for predicting emerging market credit spreads with support vector regression**

Presenter: **Gary Anderson**, CEMAR LLC, United States

Co-authors: Alena Audzeyeva

A coherent framework is proposed using support vector regression (SVR) for generating and ranking a set of high quality models for predicting emerging market sovereign credit spreads. Our framework adapts a global optimization algorithm employing a cross-validation metric for models with serially correlated variables, to produce robust sets of tuning parameters for SVR kernel functions. In contrast to previous approaches identifying a single best tuning parameter setting, we proceed with a collection of tuning parameter candidates, employing the Model Confidence Set test to select the most accurate models from the collection of promising candidates. Using bond credit spread data for three large emerging market economies and an array of input variables motivated by economic theory, we apply our framework to identify small sets of SVR models with superior out-of-sample forecasting performance. Benchmarking our SVR forecasts against random walk and conventional linear model forecasts provides evidence for the superior forecasting accuracy of SVR-based models. In contrast to linear model benchmarks, the SVR-based models can generate accurate forecasts using only the country-specific credit-spread-curve factors, lending some support to the rational expectation theory of the term structure in the context of emerging market credit spreads. Our evidence indicates a better ability of SVR to capture investor expectations about future spreads reflected in today's credit spread curve.

C1765: **Combining econometrics and machine learning to forecast realized volatility of exchange rates**

Presenter: **Aleksandr Pereverzin**, University of East Anglia, United Kingdom

Time series of financial volatility is well known for having a complex structure including several heterogeneous patterns: linear and nonlinear, long-run and short-run, etc. We propose a new two-component model of realized volatility that is based on methodological combination of econometric and machine learning approaches. In our model, Autoregressive Fractionally Integrated Moving Average (ARFIMA) framework is used to capture the linear component of realized volatility while the artificial neural network is used to model the corresponding nonlinear part. We also develop a modification of the cyclical volatility model where artificial neural networks are used to model both trend and cyclical components of realized volatility. The proposed models provide an improvement in out-of-sample forecasting accuracy over the competing approaches.

Monday 16.12.2019 14:25 - 16:05 Parallel Session O – CFE-CMStatistics

EO280 Room CLO B01 MODELLING FUNCTIONAL DATA Chair: Fabian Scheipl

E0433: **A unified framework for joint sparse clustering and alignment of functional data**
Presenter: **Valeria Vitelli**, University of Oslo, Norway

When considering functional clustering, it is often of interest to also select the portions of the domain that are most relevant to the classification purposes. However, in case the functions show the presence of misalignment, this can confound the sparse clustering procedure, possibly leading to meaningless results. The only approach currently available in this situation consists in aligning the curves first, and then using a sparse functional clustering method to estimate the groups and select the domain. However, it has been already proved that aligning and clustering the curves jointly is beneficial for the analysis. The aim is thus to jointly perform all these tasks: functional clustering, aligning the curves, and performing domain selection. The proposed method is studied in its well-posedness, and its validity is explored for a variety of measures of closeness of functional data. Indeed, the choice of the functional metric is crucial both to the purposes of functional clustering, and for the properties required to the aligning function. The method is then tested on simulated data, and its use on neuroimaging data is also explored.

E1188: **A functional additive mixed model for multivariate functional data**
Presenter: **Alexander Volkmann**, Humboldt University of Berlin, Germany
Co-authors: Almond Stoecker, Fabian Scheipl, Sonja Greven

Multivariate functional data can be intrinsically multivariate like movement trajectories in 2D or complementary like acoustics and articulation in speech production. We propose a multivariate functional additive mixed model (MFAMM) and show its application to these data situations. The approach models the dependency structure between the dimensions directly using multivariate functional principal component analysis. Multivariate functional random intercepts capture the correlation within the functions and between the multivariate functional dimensions. They also allow us to extend the model to include further between-function correlation as induced by e.g. repeated observations. The applications show that a multivariate modeling approach is more parsimonious compared to fitting independent univariate models to the data. Modeling the dependency structure between the dimensions can also generate additional insight into the properties of the multivariate functional process. A direct comparison of the multivariate and univariate approach also suggests that the estimated confidence regions might be more efficient for the MFAMM.

E1685: **Function-on-scalar quantile regression with application to mass spectrometry proteomics data**
Presenter: **Jeffrey Morris**, University of Pennsylvania, United States
Co-authors: Yusha Liu, Meng Li

Mass spectrometry proteomics, characterized by spiky, spatially heterogeneous functional data, can be used to identify potential cancer biomarkers. Existing mass spectrometry analyses utilize mean regression to detect spectral regions that are differentially expressed across groups. However, given the inter-patient heterogeneity that is a key hallmark of cancer, many biomarkers are only present at aberrant levels for a subset of, not all, cancer samples. Differences in these biomarkers can easily be missed by mean regression, but may be detected by quantile-based approaches, so we propose a quantile regression framework for functional responses. Our approach utilizes an asymmetric Laplace working likelihood, and uses basis representations and global-local shrinkage priors to enable borrowing of strength from nearby locations. A scalable Gibbs sampler is developed to generate posterior samples that can be used to perform Bayesian estimation and inference while accounting for multiple testing. Our framework performs quantile regression and coefficient regularization in a unified manner, allowing them to inform each other and leading to improvement in performance over competing methods as demonstrated by simulation studies. We apply this model to identify proteomic biomarkers of pancreatic cancer that are differentially expressed for a subset of cancer patients compared to the normal controls, which were missed by previous mean-regression based approaches.

E0846: **Functional data analysis applications to omics sciences**
Presenter: **Marzia Cremona**, Universita Laval, Canada
Co-authors: Francesca Chiaromonte, Kateryna Makova

Recent progress in sequencing technology has revolutionized the study of genomic and epigenomic processes. These assays generate massive, high-resolution Omics data often suitable to be represented as curves over the genome, whose analysis poses great challenges to standard statistical methods. Indeed, these data pose several problems such as large amounts of noise and correlations among neighboring measurements. Functional data analysis (FDA) can overcome these problems, reducing data dimension and filtering noise through smoothing, while taking advantage of the correlations among neighboring genomic positions and leveraging shape information within curves. We will present three main research directions in which FDA already proved to be effective: (1) studying and contrasting the genomic landscape of regions/loci of interests; (2) incorporating shape in the analysis of high-resolution signals of the epigenome; (3) analyzing quantitative complex phenotypes that can be represented as curves or surfaces.

EO480 Room MAL B36 THE NEW DEVELOPMENT IN THE ANALYSIS OF COMPLEX STRUCTURED DATA Chair: Wenqing He

E0317: **Modeling semi-competing risks with clusters and measurement errors**
Presenter: **Lianfen Qian**, Florida Atlantic University, United States

In lifetime data analysis, it is common to observe multiple endpoints of risks. We consider a shared frailty semi-competing model with measurement errors in covariates for cluster data with two semi-competing risks. Under the assumptions of shared Gamma frailty within each cluster and Weibull baseline hazards, we propose a corrected maximum likelihood estimation for covariate effects and Bayes estimation for the frailties. We derive the theoretical formulas for EM algorithm which is utilized for numerical optimization. To evaluate the finite sample performance of this method, we conduct the simulation studies which show that the proposed method works better than the Bayes estimation with MCMC algorithm. Moreover, the proposed method is robust to model mis-specification in terms of with or without measurement errors. For illustration purpose, we apply the proposed method to the monoclonal gammopathy of undertermined significance data. The results show that age is significant for all three baseline hazards, while the size of the monoclonal protein spike at diagnosis is significant only for the hazard from healthy to plasma cell malignancy.

E1005: **Emsembling imbalanced-spatial-structured support vector machine**
Presenter: **Xin Liu**, Shanghai University of Finance and Economics, China

The Support Vector Machine (SVM) and its extensions have been widely used in various areas due to its great prediction capability. However, these methods cannot effectively handle imbalanced data with spatial association which commonly arises from many studies such as cancer imaging study. We propose the ensembling imbalanced-spatial-structured support vector machine (EISS-SVM) method, useful for both balanced and imbalanced data. Not only does the proposed method accommodate the association between the response and the covariates, but also accounts for the spatial correlation existing in the data. Our EISS-SVM classifier embraces the usual SVM as a special case. The proposed method outperforms the competing classifiers shown in both the simulation studies and an application to real imaging data from an ongoing prostate cancer research conducted in Canada.

E1048: **Dynamic tilted current correlation for high dimensional variable screening**
Presenter: **Wenqing He**, University of Western Ontario, Canada

Variable screening is an essential procedure in high dimensional data analysis to reduce dimensionality and ensure the applicability of certain statistical methods. It is a complicated and computationally burdensome procedure since spurious correlations commonly exist among predictor variables, and important predictor variables may not have large marginal correlations with the response variable. We propose a new estimator for the correlation between the response and high-dimensional predictor variables and develop a new variable screening technique for high dimensional data based on the proposed correlation estimator. The proposed screening method enjoys the approximate sure screening and consistency properties and is capable of picking up the relevant predictor variables within a finite number of steps. It has been justified theoretically numerical through simulation and a real gene expression data.

E1759: **Simultaneous selection and inference for varying coefficients with zero regions: A soft thresholding approach**
Presenter: **Yi Li**, University of Michigan, United States

Varying coefficient models have emerged as an important tool to explore dynamic patterns in many scientific areas, such as biomedicine, finance, and epidemiology. An often overlooked aspect, however, is that some varying coefficients may have regions where the effects are zero. In a preoperative opioid use study, it was found that the association between opioid use and pain level only exists among patients with the body mass index between 25 and 30. Detection of no-effect regions of the body mass index, referred to as zero regions, is important for opioid prescription management. However, most existing methods ignore detection of zero regions. To fill this knowledge gap, we propose a new soft-thresholded varying coefficient model, where the varying coefficients are piecewise smooth with zero regions. Our new modeling approach enables us to perform variable selection and detect the zero regions of selected variables simultaneously, obtain point estimates of the varying coefficients with zero regions and construct the associated sparse confidence intervals. We prove the asymptotic properties of the estimator, and our simulation study reveals that the confidence intervals achieve the desired coverage probability. The utility of the method is further demonstrated via extensive simulation studies as well as analysis of the aforementioned preoperative opioid use study.

EO226 Room Bloomsbury RECENT DEVELOPMENTS IN STATISTICAL MODELS FOR SURVIVAL DATA Chair: Marialuisa Restaino

E0661: **Nonparametric estimation of hazard rate function from doubly truncated data under dependence**
Presenter: **Carla Moreira**, University of Minho, Portugal
Co-authors: Jacobo de Una-Alvarez

In survival analysis, the observed lifetimes often correspond to those individuals with event (infection, death and so on) occurring within a specific calendar time interval, leading to the so-called interval sampling scenario. With interval sampling, the lifetimes are doubly truncated at times determined by the birth dates and the sampling interval. Double truncation may induce a systematic bias in estimation, so specific corrections must be considered. A relevant target in survival analysis is the hazard rate function, which represents the instantaneous probability of the event of interest. We introduce a flexible estimation approach for the hazard rate under double truncation, based on kernel methods, when the lifetime and the truncation times may be dependent. The proposed estimator is constructed on the basis of a copula function which represents the dependence structure between the lifetime and the truncation times. Properties of the proposed estimator are investigated both theoretically and through simulations. Applications to the age of diagnosis of Acute Coronary Syndrome (ACS) and AIDS incubation times are performed.

E0797: **A piecewise Weibull survival model for the analysis of breast cancer subject to endogenous and exogenous factors**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca -DISMEQ, Italy
Co-authors: Juan Eloy Ruiz-Castro

Markov processes are considered in the survival literature to model the evolution of illnesses by time. In this field, the homogeneity is a usual premise but sometimes it is a very strong constraint which can lead to an erroneous modeling. In this way the non-homogeneity arises in a natural way, since the transition probabilities between two any states are not constant by time generally. A general piecewise non-homogeneous Markov model is developed and it is considered to study the evolution of breast cancer. Multiple cutpoints are considered and discrete weibull distributions are included in the model. The model is built and interesting measures calculated. The effect of treatments, menopausal state and number of infected axillary glands are incorporated as time-depending covariates. Thus, we have a multi-state model with multidimensional covariates. The results are compared with the empirical and homogeneous cases for several risk groups. Implementations in Matlab and R are provided.

E1219: **Evaluating the effect of healthcare providers through semi-Markov multi-state model and nonparametric discrete frailty**
Presenter: **Francesca Gasperoni**, MRC Biostatistics Unit, University of Cambridge, United Kingdom
Co-authors: Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples

Novel exploratory statistical methodology is introduced for investigating healthcare providers' effect on different patients outcomes through clinical administrative databases. The main purpose consists in identifying clusters of providers (latent populations) on the basis of patients' characteristics, considering the whole healthcare patients clinical history: re-admissions, discharges and death. We propose a semi-Markov multi-state model to describe the duration of hospital stay, time between hospital discharge and re-admission and time to death during admission and outside of hospital. Transition-specific hazards are modelled through a Cox proportional hazards model with a nonparametric discrete frailty term that is shared among patients hospitalised in the same provider. The inclusion of a nonparametric discrete frailty allows us to detect latent populations for each specific transition. The estimates are computed through a tailored Expectation-Maximization algorithm. As a direct consequence, we are able to identify the most frequent and most extreme latent populations across all transitions. This result is of interest for healthcare managers that can further investigate those providers associated to the most extreme latent populations' patterns. The proposed method is illustrated through an application to Heart Failure patients recorded in an administrative database from Lombardia, a northern region in Italy.

E1903: **Disease risk estimation under clustered and outcome-dependent sampling**
Presenter: **Marianne Jonker**, Radboudumc, Netherlands
Co-authors: Mar Rodriguez-Girondo

Family-based cohort studies are often used sources of data for the estimation of survival and risk prediction in populations with a high genetic predisposition to a disease. Despite the abundance of family-based data available and the increasing need in the medical world for statistical expertise to analyze these data, the development of statistical methodology is slow. Some of the challenges that are encountered when analyzing these types of data are the way the data are collected (families are ascertainment based on the presence of the disease of interest), clustering due latent factors, missing data structure (family structure) and right or interval censoring of the event of interest. An overview of existing methods is given with a focus on more recent developments, illustrated with some medical applications.

EO765 Room G21A BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS I Chair: Ines M del Puerto

E1539: **Asymptotic inference for branching random walks with immigration and applications**
Presenter: **Anand Vidyashankar**, George Mason University, United States

Branching Random Walk with Immigration (BRWI) models and their variants are useful for modeling a variety of physical, technological, biological, and financial phenomena and some of their theoretical properties have recently been studied in the literature. We focus on inference for functionals of the process; Specifically, the Laplace transform (LT) of the BRWI. We establish that there exists an interval, defined through a critical parameter, in which the LT is consistently estimable even if the BRWI process is only partially observed. Additionally, we establish the asymptotic normality of the estimator. Finally, we provide some applications to inference for the cascade generator in the study of conservative cascades.

E1528: **Branching random walks: Theoretical and simulation results**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia

The focus is on various models of a continuous-time process with generation and walking of particles on multidimensional lattices. Points of the lattice, at which the particle generation, that is birth and death of particles, can occur, are called sources of branching, and the process itself is called a branching random walk (BRW). A series of asymptotic results are provided on the behavior of the particle numbers and/or their integer moments for the following models: 1) a symmetric BRW with one source of branching and a finite or infinite number of the initial particles; 2) a symmetric BRW with a finite number of sources of various positive intensities and one initial particle; 3) a BRW with pseudo-sources, admitting possible violation of symmetry of an underlying random walk at sources of branching and one initial particle; 4) a BRW with sources at every lattice point, in which the reproduction law is described by a critical Bienamye-Galton-Watson process, and infinite number of the initial particles. In addition to the limit theorems, the simulations results based on the Monte Carlo method is represented.

E1707: **Modeling Y-linked pedigrees through branching processes**
Presenter: **Cristina Gutierrez Perez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Rodrigo Martinez Quintana

A multidimensional two-sex branching process is introduced to model the evolution of a pedigree originating from the mutation of an allele of a Y-linked gene in a monogamous population. The study of the extinction of the mutant allele and the analysis of the dominant allele in the pedigree is addressed on the basis of the classical theory of multi-type branching processes. The asymptotic behaviour of the number of couples of different types in the pedigree is also derived. Finally, using the estimates of the mean growth rates of the allele and its mutation provided by a Gibbs sampler, a real Y-linked pedigree associated with hearing loss is analysed, concluding that this mutation will persist in the population although without dominating the pedigree.

EO094 Room G3 STATISTICAL CHALLENGES IN POLICY-RELEVANT PROBLEMS Chair: Jennifer Hill

E0516: **Social network dependence, the replication crisis, and (in)valid inference**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States
Co-authors: Youjin Lee

It will be shown that social network structure can result in a new kind of structural confounding (confounding by network structure), potentially contributing to replication crises across the health and social sciences. Researchers in these fields frequently sample subjects from one or a small number of communities, schools, hospitals, etc., and while many of the limitations of such convenience samples are well-known, the issue of statistical dependence due to social network ties has not previously been addressed. A paradigmatic example of this is the Framingham Heart Study (FHS). Using a statistic that we adapted to measure network dependence, we test for network dependence and for possible confounding by network structure in several of the thousands of influential papers published using FHS data. Results suggest that some of the many decades of research on coronary heart disease, other health outcomes, and peer influence using FHS data may be biased (away from the null) and anticonservative due to unacknowledged network structure.

E1046: **Nearly exact matching in the presence of networks**
Presenter: **Alexander Volfovsky**, Duke University, United States

A classical problem in causal inference is that of matching treatment units to control units in an observational dataset. This problem is distinct from simple estimation of treatment effects as it provides additional practical interpretability of the underlying causal mechanisms that is not available without matching. Some of the main challenges in developing matching methods arise from the tension among the desire for granular and interpretable matched groups while having enough data to learn causal effects while dealing with complicating factors such as networks and non-independence among units. To deal with the influence of networks we propose to learn which network components are relevant to our causal questions. We propose several optimization objectives for match quality that capture covariates and structures that are integral for making causal statements while encouraging as many matches as possible.

E1065: **Scaling Bayesian probabilistic record linkage with post-hoc blocking**
Presenter: **Jared Murray**, University of Texas at Austin, United States

Probabilistic record linkage (PRL) is the process of determining which records in two databases correspond to the same underlying entity in the absence of a unique identifier. Bayesian solutions to this problem provide a powerful mechanism for propagating uncertainty due to uncertain links between records (via the posterior distribution). However, computational considerations severely limit the practical applicability of existing Bayesian approaches. We propose a new computational approach yielding a restricted MCMC algorithm that samples from an approximate posterior distribution. Our advances make it possible to efficiently perform Bayesian PRL for large problems. We demonstrate the methods on a subset of an OCR'd dataset, the California Great Registers, a collection of 57 million voter registrations from 1900 to 1968 that comprise the only panel data set of party registration collected before the advent of scientific surveys.

E1076: **Model interpretation through lower dimensional posterior summarization**
Presenter: **Carlos Carvalho**, The University of Texas at Austin, United States
Co-authors: Jared Murray

Nonparametric regression models have recently surged in their power and popularity, accompanying the trend of increasing dataset size and complexity. While these models have proven their predictive ability in empirical settings, they are often difficult to interpret and do not address the underlying inferential goals of the analyst or decision maker. We propose a modular two-stage approach for creating parsimonious, interpretable summaries of complex models which allow freedom in the choice of modeling technique and the inferential target. In the first stage a flexible model is fit which is believed to be as accurate as possible. In the second stage, lower-dimensional summaries are constructed by projecting draws from the distribution onto simpler structures. These summaries naturally come with valid Bayesian uncertainty estimates. Further, since we use the data only once to move from prior to posterior, these uncertainty estimates remain valid across multiple summaries and after iteratively refining a summary. We explore extension do causal non-parametric models.

EO548 Room G5 FRONTS AND FRONTIERS: RECENT STUDIES IN MODELING AND ESTIMATION Chair: Carlos Martins-Filho

E0604: **On the estimation and treatment of observation uncertainty in data assimilation for numerical weather prediction**
Presenter: **Sarah Dance**, University of Reading, United Kingdom

In numerical weather prediction, forecasts are produced each hour via the numerical solution of a partial differential equation, starting from an initial estimate of the current state of the atmosphere. Variational data assimilation is used to produce these initial data, combining the latest observational data (of order 10^7 observations) with numerical model forecasts (with state vector of dimension around 10^9) to estimate the current state of the system. The variational assimilation problem has a Bayesian formulation, in which the observation and model forecast are assumed to be Gaussian variables with prescribed covariances. Until recently, most operational forecasting centres have assumed that the errors in the observations are uncorrelated. However, this is not always true, especially when considering multichannel satellite observations, and it has been shown that fully specifying observation error covariance matrices leads to more accurate forecasts. The latest approaches used in numerical weather

prediction to estimate and treat observation uncertainty are reviewed. We will discuss covariance estimation, regularization and implementation in high performance computers.

E0989: **Model selection for migrating cell fronts**
Presenter: **David Bortz**, University of Colorado - Boulder, United States
The collective migration of cells is a central biological process for multicellular organisms. A central obstacle inhibiting deeper study of migration lies in the fact that there are large number of viable explanatory mechanisms. We will discuss the development and limitations of using information-theoretic model selection criteria for dynamical systems. The ultimate goal is to be able to choose among several candidate models to find those which best represent the migration patterns - and thus those that best describe the underlying biology.

E0611: **Bayesian inference in structured epidemics**
Presenter: **Vanja Dukic**, University of Colorado at Boulder, United States
The purpose is to discuss Bayesian state space modeling and inference suitable for on-line epidemic surveillance, using flu as an example.

E1262: **Robust estimation of additive boundaries with quantile regression and shape constraints**
Presenter: **Carlos Martins-Filho**, University of Colorado at Boulder, United States
Co-authors: Lan Xue, Lijian Yang, Yan Fang
The estimation of the boundary of a set is considered when it is known to be sufficiently smooth, to satisfy certain shape constraints and to have an additive structure. The proposed method is based on spline estimation of a conditional quantile regression and is resistant to outliers and/or extreme values in the data. It extends previous work and can also be viewed as an alternative to existing estimators that have been widely used in empirical analysis. The results of a Monte Carlo study show that the new method significantly outperforms the commonly used methods when outliers or heterogeneity are present. The theoretical analysis indicates that our proposed boundary estimator is uniformly consistent under a set of standard assumptions. We illustrate practical use of our method by estimating two production functions using real-world data sets.

EO843 Room Gordon BAYESIAN DESIGN OF EXPERIMENTS Chair: Markus Hainy

E0378: **Bayesian optimal design for ordinary differential equation models**
Presenter: **Antony Overstall**, University of Southampton, United Kingdom
Bayesian optimal design is considered for physical models derived from the (intractable) solution to a system of ordinary differential equations (ODEs). Bayesian optimal design requires the minimisation of the expectation (over all unknown and unobserved quantities) of an appropriately chosen loss function. This can be non-trivial due to 1) the high dimensionality of the design space; and 2) the intractability of the expected loss. In this case, a further complication arises from the intractability of the solution to the system of ODEs. We propose a strategy that employs a modification of the continuous coordinate exchange algorithm where a statistical emulator is employed to approximate the expected loss function, and a probabilistic solution to the system of ODEs. The strategy is demonstrated on several illustrative examples from the biological sciences.

E0530: **Simulation-based optimal sequential Bayesian design using policy gradient reinforcement learning**
Presenter: **Xun Huan**, University of Michigan, United States
Co-authors: Wanggang Shen
Experiments are indispensable for learning and developing models in science and engineering. When experiments are expensive, a careful design of these limited data-acquisition opportunities can be immensely beneficial. Optimal experimental design, while leveraging the predictive capabilities of a simulation model, provides a rigorous framework to systematically quantify and maximize the value of an experiment. We focus on the design of a finite sequence of experiments, seeking design policies (strategies) that can (a) adapt to newly collected data during the sequence (i.e. feedback) and (b) anticipate future changes (i.e. lookahead). We cast this sequential learning problem in a Bayesian setting with information-based utilities, and solve it numerically via policy gradient methods from reinforcement learning. In particular, we directly parameterize the policies and value functions—thus adopting an actor-critic approach—and improve them using gradient estimates produced from simulated design sequences. The overall method is demonstrated on an algebraic benchmark and a sensor placement application for source inversion. The results provide intuitive insights on the benefits of feedback and lookahead, and indicate substantial computational advantages compared to previous numerical methods based on approximate dynamical programming.

E1069: **Optimal design for infectious diseases**
Presenter: **David Price**, The University of Melbourne, Australia
Observational studies to understand infectious diseases are often conducted in resource-limited settings. Additionally, models of infectious disease dynamics are often complex. We will discuss some methods for evaluating the utility of a design and searching across the design space, that are particularly beneficial when considering designs for infectious disease dynamics. We will provide some examples of models of infectious diseases where we have used these methods to evaluate and find optimal designs.

E1192: **Optimal Bayesian design for models with intractable likelihoods via supervised learning methods**
Presenter: **Markus Hainy**, Johannes Kepler University, Austria
Co-authors: David Price, Olivier Restif, Christopher Drovandi
Optimal Bayesian experimental design is often computationally intensive due to the need to approximate many posterior distributions for datasets simulated from the prior predictive distribution. The issues are compounded further when the statistical models of interest do not possess tractable likelihood functions and only simulation is feasible. We employ supervised learning methods to facilitate the computation of utility values in optimal Bayesian design. This approach requires considerably fewer simulations from the candidate models than previous approaches using approximate Bayesian computation. The approach is particularly useful in the presence of models with intractable likelihoods but can also provide computational advantages when the likelihoods are manageable. We consider the two experimental goals of model discrimination and parameter estimation. The methods are applied to find optimal designs for models in epidemiology and cell biology.

EO851 Room MAL G13 EMPIRICAL BAYES IN THE 21ST CENTURY Chair: Roger Koenker

E1552: **Nonparametric empirical Bayes methods for sparse, noisy signals**
Presenter: **Junhui Jeffrey Cai**, University of Pennsylvania, United States
Co-authors: Linda Zhao
High dimensional signal recovering problems will be considered. The goal is to identify the true signals from the noise controlling false discovery rate and then to make inference for the unknown signals. We propose a nonparametric empirical Bayesian scheme to tackle the problem. The method adapts well to varying degrees of sparsity. It not only performs well to recover the signals, but also provides credible intervals. The method is built upon noisy data with exponential family distribution. It covers large range of data structure such as normal means with heteroskedastic variance, Poisson data with varying degrees of frequency, and Binomial counts. Simulations show that our method outperforms existing ones. Applications in microarray data as well as sport data such as predicting batting averages will be discussed.

E1127: **A regression modeling perspective on compound decision problems**
Presenter: **Sihai Zhao**, University of Illinois at Urbana-Champaign, United States

Co-authors: William Biscarri

Compound decision theory is a classical area of statistics that is now experiencing a resurgence of interest. Current work is dominated by empirical Bayes approaches with desirable theoretical and empirical properties. However, empirical Bayes methods are at odds with both frequentist and Bayesian philosophies, and furthermore are not flexible enough to accommodate more complicated problems. We will present a new regression modeling perspective on compound decision problems, expanding upon previous work, that interprets the James-Stein estimator as a linear regression estimator. This new perspective will motivate new flexible estimation methods that can easily incorporate auxiliary information and new inferential procedures for the resulting estimators. These new tools will be illustrated in the analysis of genomic data.

E1123: **Nonparametric maximum likelihood estimation of mixture models: Recent developments**

Presenter: **Ivan Mizera**, University of Alberta, Canada

Co-authors: Sile Tao

The purpose is to review primal and dual formulations for the nonparametric maximum likelihood estimation of the mixing distribution in mixture models, in the context of the empirical Bayes methodology known as the Kiefer-Wolfowitz, or also Robbins method. While the original, primal formulation, is an infinite-dimensional convex optimization problem, its dual has finite-dimensional objective function and infinite-dimensional constraint. This opens room for possible alternative strategies, with an eye of their potential scalability to high-dimensional problems, which are subjected to some theoretical analysis, and also tried on practical examples.

E1104: **Covariate-powered empirical Bayes estimation**

Presenter: **Nikolaos Ignatiadis**, Stanford University, United States

Co-authors: Stefan Wager

How to simultaneously analyze many noisy experiments in the presence of rich covariate information? The goal of the analyst is to optimally estimate the true effect underlying each experiment. Both the noisy experimental results and the auxiliary covariates are useful for this purpose, but neither data source on its own captures all the information available to the analyst. A flexible plug-in empirical Bayes estimator is proposed that synthesizes both sources of information and may leverage any black-box predictive model. The approach is within a constant factor of minimax for a simple data-generating model. Furthermore, an extension to the classic result of James-Stein is established, whereby the proposed estimator dominates the sample mean of the experimental results under quadratic risk; even if the auxiliary covariates contain no information about the true effects. Finally, the method exhibits promising empirical performance on both real and simulated data.

EO122 Room MAL G14 PROJECTION PURSUIT: APPLICATIONS Chair: Nicola Loperfido

E0241: **Higher order moments of the estimated tangency portfolio weights**

Presenter: **Stepan Mazur**, Orebro University, Sweden

Co-authors: Farrukh Javed, Edward Ngailo

The estimated weights of the tangency portfolio are considered. We derive analytical expressions for the higher order non-central and central moments of these weights when the returns are assumed to be independently and multivariate normally distributed. Moreover, the expressions for mean, variance, skewness and kurtosis of the estimated weights are obtained in closed-forms. Later, we complement our results with a simulation study where data from the multivariate normal and t-distributions are simulated and the first four moments of estimated weights are computed by using the Monte Carlo experiment. It is noteworthy to mention that the distributional assumption of returns is found to be important, especially, for the first two moments. Finally, through an empirical study utilizing returns of four financial indices listed in the NASDAQ stock exchange, we observe the presence of time dynamics in higher moments.

E0265: **Fourth cumulant for the random sum of random vectors**

Presenter: **Farrukh Javed**, Orebro University, Sweden

Co-authors: Nicola Loperfido, Stepan Mazur

The fourth cumulant for the aggregated multivariate claims is considered. A formula is presented for the general case when the aggregating variable is independent of the multivariate claims. Two important special cases are considered. In the first one, multivariate skewed normal claims are considered and aggregated by a Poisson variable. For the Poisson Skew-normal case, we also proved that the kurtosis of a linear projection of aggregated multivariate claims attains its maximum when the projecting direction is the same as the shape parameter α. The second case is dealing with multivariate asymmetric generalized Laplace and aggregation is made by a negative binomial variable. Due to the invariance property, the latter case can be derived directly, leading to the identity involving the cumulant of the claims and the aggregated claims. There is a well-established relationship between asymmetric Laplace motion and negative binomial process that corresponds to the invariance principle of the aggregating claims for the generalized asymmetric Laplace distribution. We explore this relationship and provide multivariate continuous time version of the results. It is discussed how these results that deal only with dependence in the claim sizes can be used to obtain a formula for the fourth cumulant for more complex aggregate models of multivariate claims in which the dependence is also in the aggregating variables.

E0363: **Multivariate kurtosis with the R package MultiKurt**

Presenter: **Cinzia Franceschini**, Tuscia University, Italy

Co-authors: Nicola Loperfido

MultiKurt is an R package purported to describe, testing and visualize multivariate kurtosis. In particular, it incorporates state-of-the-art algorithms for computing linear projections which either maximize, minimize or remove kurtosis. MultiKurt also computes scalar-valued and matrix-valued measures of multivariate kurtosis. The usage of MultiKurt is illustrated with well-known data sets.

E1788: **An improved method of combining forecasts based on fourth cumulant**

Presenter: **Massimiliano Giacalone**, University of Naples - Federico II, Italy

Co-authors: Raffaele Mattera

A well-known result in statistics and econometrics is that a linear combination of two point forecasts has a smaller Mean Square Error (MSE) than the two competing forecasts themselves. The kind of combination methods are various, ranging from the simple average (SA) to more robust methods as the one based on median or on a Trimmed Average (TA) to other methods based on regression or optimization techniques. Using the regression-based approach, the resulting combined forecast is a linear function of the individual forecasts where the weights are estimated via Ordinary Least Squares (OLS), minimizing the sum of squared errors. A clear advantage of the OLS forecast combination method is that the combined resulting forecast is unbiased even if one of the individual forecasts is biased. Other alternative methods were developed, implementing the minimization of a different loss function, as happen with the least absolute sum of squares. However, these methods may fail to get a realistic result if the forecasts density are heavy-tailed, as happen in many situation (e.g. financial time series). Therefore, we propose a forecast combination method based on Lp-norm estimator, where the minimization of residuals is done according to estimated data kurtosis and the selection of more relevant forecast is achieved via a projection pursuit based on fourth cumulant. A simulation study is presented in order to show improvements in forecasting accuracy.

EO352 Room MAL G15 TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS III Chair: Raffaele Argiento

E1070: **A Bayesian nonparametric approach for functional regression with application to sport data**
Presenter: **Alessandro Lanteri**, University of Turin, Italy
Co-authors: Raffaele Argiento, Silvia Montagna, James Hopker

In sport analytics, there is often interest in predicting elite athletes performance at a future sporting event given his/her competitive results tracked throughout the athlete's career and other (time-varying) covariates. Such predictions can be useful both for scouting purposes, and to build red flag indicators of unexpected increases in athlete performance for targeted anti-doping testing. We propose a predictive model for the longitudinal trajectory of athletes performance where we characterize the curve with a sparse basis expansion allowing individual time-dependant covariates to impact the shape of the estimated trajectories. Moreover, we introduce random intercepts, distributed according to a nonparametric hierarchical process, in order to induce clustering while borrowing statistical information across curves. In particular, we assume a hierarchical normalized generalized gamma process to grants great flexibility in clustering and accuracy in prediction. We apply our model to a longitudinal study on shot put athletes, where their competitive results are tracked throughout their career.

E1296: **Bayesian divisive clustering**
Presenter: **Paul Kirk**, University of Cambridge, United Kingdom
Co-authors: Christopher Foley

A novel model-based Bayesian divisive clustering algorithm is presented. The algorithm starts with all observations in the same cluster, and iteratively divides the cluster into sub-clusters. Whether or not a cluster should be subdivided is determined using a Bayesian model selection approach, based on the calculation of (approximate) Bayes factors. Adopting an appropriate choice of prior on the space of partitions, we establish links to the Dirichlet process mixture model, and derive approximations that vastly improve scalability. We provide a case study application from genetics, in which traits are clustered together if they share a common causal variant. We demonstrate that the scalability of our approach enables us to perform analyses that would be impossible using competing state-of-the-art techniques.

E1416: **Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data**
Presenter: **Xenia Miscouridou**, Oxford, United Kingdom
Co-authors: Francois Caron, Yee Whye Teh

A novel class of network models for temporal dyadic interaction data is proposed. The goal is to capture a number of important features often observed in social interactions: sparsity, degree heterogeneity, community structure and reciprocity. We propose a family of models based on self-exciting Hawkes point processes in which events depend on the history of the process. The key component is the conditional intensity function of the Hawkes Process, which captures the fact that interactions may arise as a response to past interactions (reciprocity), or due to shared interests between individuals (community structure). In order to capture the sparsity and degree heterogeneity, the base (non time dependent) part of the intensity function builds on compound random measures. We conduct experiments on a variety of real-world temporal interaction data and show that the proposed model outperforms many competing approaches for link prediction, and leads to interpretable parameters.

E1146: **A Bayesian nonparametric approach to factor analysis**
Presenter: **Remi Piatek**, University of Copenhagen, Denmark

A new approach is introduced for the inference of non-Gaussian factor models based on Bayesian nonparametric methods. It relaxes the usual normality assumption on the latent factors, widely used in practice, which is too restrictive in many settings. The approach, on the contrary, does not impose any particular assumptions on the shape of the distribution of the factors, but still secures the basic requirements for the identification of the model. We design a new sampling scheme based on marginal data augmentation for the inference of mixtures of normals with location and scale restrictions. This approach is augmented by the use of a retrospective sampler, to allow for the inference of a constrained Dirichlet process mixture model for the distribution of the latent factors. We carry out a simulation study to illustrate the methodology and demonstrate its benefits. The sampler is very efficient in recovering the distribution of the factors, and only generates models that fulfill the identification requirements. A real data example illustrates the applicability of the approach.

EO262 Room MAL G16 MARKOV CHAIN MONTE CARLO FOR COMPLEX DATA Chair: Galin Jones

E0354: **New visualizations for Monte Carlo simulations**
Presenter: **James Flegal**, University of California - Riverside, United States
Co-authors: Nathan Robertson, Galin Jones, Dootika Vats

In Monte Carlo simulations, samples are obtained from a target distribution in order to estimate various features. We present a flexible class of visualizations for assessing the quality of estimation, which are principled, practical, and easy to implement. To this end, we establish joint asymptotic normality for any collection of means and quantiles. Using the limit distribution, we construct $1-\alpha$ level simultaneous confidence intervals, which we integrate within visualization plots. We demonstrate the utility of our visualizations in various Monte Carlo simulation settings including Monte Carlo estimation of expectations and quantiles, Monte Carlo simulation studies, and Bayesian analyses using Markov chain Monte Carlo sampling. The marginal-friendly interpretation enables practitioners to visualize simultaneous uncertainty, a substantial improvement from current visualizations.

E0519: **Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions**
Presenter: **Karl Oskar Ekvall**, Vienna University of Technology, Austria
Co-authors: Galin Jones

The use of Markov chain Monte Carlo (MCMC) to explore posterior distributions is widespread in Bayesian statistics. In order to assess or ensure the reliability of an analysis using MCMC it is essential to understand some convergence properties of the chain in use. We discuss a collapsed Gibbs sampler for Bayesian vector autoregressions with predictors, or exogenous variables. The emphasis is on how the algorithm's convergence rate is affected as the length of the sample path from the underlying vector autoregression increases. The main result, which is among the first of its kind for practically relevant MCMC algorithms, establishes an asymptotic upper bound on the convergence rate.

E0556: **Counterexamples for optimal scaling of Metropolis-Hastings chains with rough target densities**
Presenter: **Jure Vogrinc**, University of Warwick, United Kingdom
Co-authors: Wilfrid Kendall

For sufficiently smooth targets of product form it is known that the variance of a single coordinate of the proposal in RWM (Random walk Metropolis) and MALA (Metropolis adjusted Langevin algorithm) should optimally scale as n^{-1} and as $n^{-1/3}$ with dimension n, and that the acceptance rates should be tuned to 0.234 and 0.574. We establish counterexamples to demonstrate that smoothness assumptions such as having a continuous derivative for RWM and having three continuous derivatives for MALA are indeed required if these guidelines are to hold. The counterexamples identify classes of marginal targets, obtained by perturbing a standard Normal density at the level of the potential (or second derivative of the potential for MALA) by a path of fractional Brownian motion with Hurst exponent H, for which these guidelines are violated. For such targets there is strong evidence that RWM and MALA proposal variances should optimally be scaled as $n^{-1/H}$ and as $n^{-1/(2+H)}$ and will then obey anomalous acceptance rate guidelines. We will briefly discuss useful heuristics resulting from this theory.

E0781: **Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression**
Presenter: **Qian Qin**, University of Minnesota, United States
Co-authors: James Hobert
The use of MCMC algorithms in high dimensional Bayesian problems has become routine. This has spurred so-called convergence complexity analysis, the goal of which is to ascertain how the convergence rate of a Monte Carlo Markov chain scales with sample size, n, and/or number of covariates, p. The convergence complexity of Albert and Chib's algorithm for Bayesian probit regression is studied. By constructing convergence bounds with respect to some Wasserstein distance, it is found that, under reasonable data structures, the algorithm converges rapidly even when n and p are large.

EO787 Room Montague BFF: FOUNDATIONS OF STATISTICS AND THEIR IMPACTS ON APPLICATIONS Chair: Jan Hannig

E0181: **BFF: Bayesian, fiducial, frequentist analysis of age effects in daily diary data**
Presenter: **Shevaun Neupert**, North Carolina State University, United States
Co-authors: Jan Hannig
Age effects in within-person slopes in daily diary data were examined with Bayesian, Generalized Fiducial Inference (GFI), and frequentist paradigms. Daily stressor exposure data across six domains were used to generate within-person emotional reactivity slopes with daily negative affect. Systematic age differences and similarities in these reactivity slopes were tested, which are inconsistent in previous research. 116 older (aged 60-90) and 107 younger (aged 18-36) adults from the Mindfulness and Anticipatory Coping Everyday study responded to daily stressor and negative affect questions each day for eight consecutive days, resulting in 1,627 total days. Daily stressor domains included arguments, potential/avoided arguments, work/volunteer stressors, home stressors, network stressors, and health-related stressors. Using Bayesian, GFI, and frequentist paradigms, results for each of the six stressor domains with a focus on interpreting age effects in within-person reactivity were compared. As an example, frequentist multilevel models and Bayesian models with main effects of arguments (reactivity slope), age, and their interaction predicting negative affect suggested no age differences in reactivity. However, the GFI solution suggested that older adults were less reactive than younger adults. GFI is a useful tool that provides additional information when making determinations regarding null age effects in within-person slopes.

E0210: **Dempster-Shafer adaptive clinical trials and A/B tests**
Presenter: **Paul Edlefsen**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Raabya Rossenkhan
Dempster-Shafer is a statistical framework that extends Bayesian analysis by incorporating a set-valued representation of probabilities, yielding "imprecise probability" measures called Dempster-Shafer PQR values. These are posterior assessments of a hypothesis that allow for a new dimension of uncertainty: "don't know" ($R > 0$). In this framework, the outcome of an analysis might be "there is insufficient evidence to make a conclusion; collect more data". Recent advances in employing binomial Dempster-Shafer inferences sequentially in an "online" adaptive experiment are presented, where decisions to draw a conclusion or to gather further data are made sequentially. We demonstrate the power of the Dempster-Shafer adaptive binomial clinical trial design through simulation studies under various cost (loss) function scenarios, with applications to early-phase clinical testing of (for example) HIV vaccines, as well as to A/B testing in the (other sense of the) "online" setting.

E1584: **Safe testing & Probabilities: A unified treatment of optional stopping, untrustworthy priors and misspecification**
Presenter: **Peter Grunwald**, CWI and Leiden University, Netherlands
P-values give rise to valid Type-I error probabilities if sampling plan and significance level are chosen independently of the observed data, and give un-interpretable numbers in other situations such as with optional stopping. We introduce S-values, an alternative to p-values which remain valid, under specific loss functions, if the significance level may depend on the data. We review test martingales, which even remain valid under optional stopping. We introduce 'safety' as an analogue for 'validity' for methods such as Bayes and fiducial that output a distribution. This leads to a general calculus of 'validity' and 'safety' that should give us a much better idea of what possibly misspecified models can be used for and what not. Some examples: generalized Bayesian inference with a (high-dimensional) linear regression model, even if severely misspecified, is safe for squared error prediction and for assessing the quality of those predictions; it is unsafe for just about any other loss function. The fiducial posterior is safe for making confidence statements when the sampling plan is fixed in advance, but not under optional stopping. Bayesian null hypothesis testing with point null is safe for Type 0 Error Probabilities, irrespective of the prior, even under optional stopping; with composite hypothesis testing it is not safe for Type 0 Error, not even with a fixed stopping time.

E1993: **The EAS approach for graphical selection consistency in vector autoregression models**
Presenter: **Jonathan Williams**, North Carolina State University, United States
As evidenced by various recent and significant papers within the frequentist literature, along with numerous applications in macroeconomics, genomics, and neuroscience, there continues to be substantial interest to understand the theoretical estimation properties of high-dimensional vector autoregression (VAR) models. To date, however, while Bayesian VAR (BVAR) models have been developed and studied empirically (primarily in the econometrics literature) there exist very few theoretical investigations of the repeated sampling properties for BVAR models in the literature. In this direction, we construct methodology via the epsilon-admissible subsets (EAS) approach for posterior-like inference based on a generalized fiducial distribution of relative model probabilities over all sets of active/inactive components (graphs) of the VAR transition matrix. We provide a mathematical proof of pairwise and strong graphical selection consistency for the EAS approach for stable VAR(1) models which is robust to model misspecification, and demonstrate numerically that it is an effective strategy in high-dimensional settings.

EO140 Room CLO 101 STATISTICAL MODELLING, COMPARISONS, LEARNING AND DISCOVERIES Chair: Subir Ghosh

E0662: **Statistical methods for dynamic cardiovascular risk prediction**
Presenter: **Jessica Barrett**, MRC Biostatistics Unit, United Kingdom
Co-authors: Angela Wood, Michael Sweeting, Ellie Paige, David Stevens
A risk prediction model aims to accurately predict the probability of some event occurring within a pre-specified time window for a new individual. A dynamic risk prediction model allows risk predictions to be updated over time in response to new information becoming available. Methods for dynamic risk prediction include (i) using the last-observation-carried forward (LOCF) of each risk factor as a time-varying covariate in a time-to-event model, (ii) landmarking, where a discrete set of landmark times is specified at which risk predictions are to be made and survival is modelled from the landmark time only for individuals still at risk, and (iii) joint modelling, where repeated risk factor measurements and the time to event are modelled simultaneously, e.g., for cardiovascular disease (CVD), the 10-year risk of a CVD event is typically used to make clinical decisions about whether to prescribe lipid-lowering medication. Time-varying CVD risk factors, such as blood pressure, cholesterol and smoking status, may be monitored over time and used to dynamically update CVD risk predictions. Dynamic risk prediction models for CVD will be compared in different data scenarios, including a single cohort study, an individual participant data meta-analysis and UK electronic health records.

E0818: **Random effects dynamic panel models for unequally-spaced repeated measures**
Presenter: **Fiona Steele**, London School of Economics, United Kingdom
Dynamic models, also known as autoregressive or lagged response models, are widely used for the analysis of longitudinal data in social science and health. However, standard discrete-time models assume that measurements of the response and time-varying covariates are taken at the same

equally-spaced occasions. Unequal spacing is a common feature of longitudinal studies, which may arise by design or because of nonresponse. A general random effects dynamic model is proposed to handle unequally-spaced responses that are measured less frequently than time-varying covariates. The approach is suitable for continuous, binary or ordinal multivariate responses. The methodology is assessed in a simulation study, and applied to bivariate binary data on bidirectional exchanges of support between adult children and their non-coresident parents from the British Household Panel Survey and UK Household Longitudinal Study. Of particular interest are the effects of changes in children's circumstances on help received from and given to their parent(s). Using annual data on partnership and employment status and children, we estimate the effects of partnership and employment transitions between year $t-1$ and t and of the presence and age of children at t on exchanges in each direction at t. A bivariate model is used to estimate the reciprocity of exchanges.

E0913: **Bayesian optimal design of experiments motivated by challenges from science and technology**
Presenter: **David Woods**, University of Southampton, United Kingdom

The design of any experiment is implicitly Bayesian, with prior knowledge being used informally to aid decisions such as which factors to vary and the choice of plausible causal relationships between the factors and measured responses. Adoption of formal Bayesian methods allow uncertainty in these decisions to be incorporated into design selection through prior distributions that encapsulate information available from scientific knowledge or previous experimentation. Further, a design may be explicitly tailored to the aim of the experiment through a decision-theoretic approach with an appropriate loss function. However, finding decision-theoretic optimal designs is challenging, largely due to the typically high-dimensional and intractable integration required to evaluate the expected loss. We review some of the recent research in this area, with particular reference to problems motivated by experiments from science, industry and technology.

E1158: **The harmonic mean p-value for combining dependent tests**
Presenter: **Daniel Wilson**, University of Oxford, United Kingdom

Analysis of big data frequently involves statistical comparison of millions of competing hypotheses to discover hidden processes underlying observed patterns of data, for example, in the search for genetic determinants of disease in genome-wide association studies (GWAS). Controlling the familywise error rate (FWER) is considered the strongest protection against false positives, but makes it difficult to reach the multiple testing-corrected significance threshold. We discuss the harmonic mean p-value (HMP), which controls the FWER while greatly improving statistical power by combining dependent tests using generalized central limit theorem. We show that the HMP effortlessly combines information to detect statistically significant signals among groups of individually nonsignificant hypotheses in examples of a human GWAS for neuroticism and a joint humanpathogen GWAS for hepatitis C viral load. The HMP simultaneously tests all ways to group hypotheses, allowing the smallest groups of hypotheses that retain significance to be sought. The power of the HMP to detect significant hypothesis groups rivals the power of the Benjamini-Hochberg procedure to detect significant hypotheses, although the latter only controls the weaker false discovery rate (FDR). The HMP has broad implications for the analysis of large datasets, because it enhances the potential for scientific discovery.

EO544 Room Court SPECIAL PROBLEMS IN CLASSIFICATION OF DISTRESSED COMPANIES Chair: Alessandro Beretta

E0843: **Spatial modelling using clustering**
Presenter: **Jonathan Ansell**, The University of Edinburgh, United Kingdom
Co-authors: Antonia Gieschen, Raffaella Calabrese, Belen Martin-Barragan, Galina Andreeva

The consequences of recent work considering clustering and prediction of performance are addressed. These have both explore the spatial dependency within the domain of application. A paper dealt with the spatio-temporal clustering to explore General Practitioners (GPs) prescription behaviour across Scotland. It was based on National Health Service (NHS) Scotland Open Data which details prescription behaviour of GP. It employed ST-DBSCAN. A second paper investigated the impact of spatial behaviour in relation to performance of Small and Medium Sized Enterprises (SMEs) in the Greater London Area. The focus is on the development of appropriate measures of closeness expressed in the W matrix. The data covered a 4 year period and was modelled using a spatial probit model. Obviously, the link between the two analysis is expression of closeness across both spatial and other metrics of the individual subjects. This forms the central theme of the paper which will be discussed.

E0955: **Competing risks PH cure model and GEV regression: Analysis of bank failures and acquisitions in the U.S.**
Presenter: **Alessandro Beretta**, HEC Liege, Belgium
Co-authors: Cedric Heuchenne, Marialuisa Restaino

The factors influencing the disappearance of commercial banks in the United States are investigated. A bank may cease to exist primarily due to failure or acquisition by another entity and, moreover, it may not be susceptible to one or both events. For this reason, we use a competing risks proportional-hazards cure model in order to measure the impact of bank-specific and macroeconomic variables on the probabilities to experience these events (i.e. incidence) and on the survival time of susceptible banks (i.e. latency). We propose to model the incidence distribution using Generalized Extreme Value regression and compare the results with the usual logistic regression model. The proposed methodology is evaluated by means of a simulation study and then applied to a dataset of more than 4000 United States commercial banks spanning the period 1993-2018.

E1175: **Contagion effects for UK small business failures: A spatial hierarchical autoregressive model for binary data**
Presenter: **Raffaella Calabrese**, University of Edinburgh, United Kingdom

The focus is on modelling contagion effects between and within groups on small business failures in London. Small business clusters could be defined based on different companies' characteristics, for example economic sector or geographical location. These aspects are usually included as fixed effects to predict the defaults of small and medium-sized enterprises (SMEs). This approach however ignoresthe interaction between the companies' groups and captures only the heterogeneity across the clusters. To include both contagion effects between and within groups, we propose a Bayesian hierarchical model for binary data. We find that the contagion component at the lower level, based on the geographical location, is not significant if we ignore the clustering. However, it becomes significant if we consider the industry group effect and also the upper level interdependence is significant. Finally, we show that our proposal improves the ability to predict SMEs defaults in London.

E1363: **Investing in high-yield debt: The case of U.S. bond market**
Presenter: **Thomas Aeschbacher**, University of St Gallen, Switzerland
Co-authors: Alexander Kostrov

In the era of shrinking stock market returns, bond market attracts much investor's attention. We apply machine learning techniques to classify bond issues as distressed in the U.S. corporate bond market. There is a large-scale data preparation exercise behind our analysis. We describe some peculiarities in investor's behavior and try to exploit them in order to enhance investment strategies in the high-yield market segment. It is shown that improved prediction accuracy of bond defaults in statistical terms leads to higher economic gain for an investor.

EO482 Room MAL 152 SEMIPARAMETRIC METHODS FOR RISK EVALUATION Chair: Yu Cheng

E0332: **Causal proportional hazards regression with a binary instrumental variable**
Presenter: **Limin Peng**, Emory University, United States
Co-authors: Behzad Kianian, Jung Kim, Jason Fine

Instrumental variables (IV) are a useful tool for estimating causal effects in the presence of unmeasured confounding. IV methods are well developed for uncensored outcomes, particularly for structural linear equation models, where simple two-stage estimation schemes are available.

The extension of these methods to survival settings is challenging, partly because of the nonlinearity of the popular survival regression models and partly because of the complications associated with right censoring or other survival features. Motivated by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer screening trial, we develop a simple causal hazard ratio estimator in a proportional hazards model with right censored data. The method exploits a special characterization of IV which enables the use of an intuitive inverse weighting scheme that is generally applicable to more complex survival settings with left truncation, competing risks, or recurrent events. We rigorously establish the asymptotic properties of the estimators, and provide plug-in variance estimators. The proposed method can be implemented in standard software, and is evaluated through extensive simulation studies. We apply the proposed IV method to a data set from the Prostate, Lung, Colorectal and Ovarian cancer screening trial to delineate the causal effect of flexible sigmoidoscopy screening on colorectal cancer survival which may be confounded by informative noncompliance with the assigned screening regimen.

E1245: **Quantile regression on life lost**
Presenter: **Jong-Hyeon Jeong**, University of Pittsburgh, United States

The lost lifespan concerns time from occurrence of an event of interest to the current time point and has recently emerged as a new summary measure for cumulative information inherent in time-to-event data. This summary measure provides several benefits over the traditional methods, including more straightforward interpretation yet less sensitivity to heavy censoring. We propose a regression method for the quantiles of the lost lifespan distribution under right censoring. The consistency and asymptotic normality of the regression parameters are discussed. We propose a computationally efficient method for estimating the variance-covariance matrix of the regression coefficient estimates. Simulation results are presented to validate the finite sample properties of the proposed estimators and test statistics. The proposed method is illustrated with a real dataset from a breast cancer study.

E1112: **Ensemble estimation and variable selection with semiparametric transformation models**
Presenter: **Sunyoung Shin**, University of Texas at Dallas, United States

Semiparametric transformation models associate potentially time-dependent covariates on survival time. We consider a certain class of semiparametric transformation models, whose likelihood factors into separate components. When an efficient estimator of the regression parameter is available for each component, an optimal weighted combination of the component estimators, named an ensemble estimator, may be employed as an overall estimate of the regression parameter. This approach is useful when the full likelihood function may be difficult to maximize but the components are easy to maximize. Variable selection is important in such regression modelling but the applicability of existing techniques is unclear in the ensemble approach. We propose ensemble variable selection using the least squares approximation technique on the unpenalized ensemble estimator, followed by ensemble re-estimation under the selected model. We conduct numerical studies with proportional odds models to show that the proposed method outperforms alternative approaches.

E0612: **Novel metrics for assessing importance of new biomarkers for competing outcomes**
Presenter: **Yu Cheng**, University of Pittsburgh, United States
Co-authors: Zheng Wang, Eric Seaberg, James Becker

The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were originally proposed to characterize accuracy improvement in predicting a binary outcome, when new biomarkers are added to regression models. These two indices have been extended from dichotomous outcomes to multi-categorical and survival outcomes. Working on an AIDS study where the onset of cognitive impairment is competing risks censored by death, we extend the NRI and the IDI to competing risk outcomes, by using cumulative incidence functions to quantify cumulative risks of competing events, and adopting the definitions of the two indices for multi-category outcomes. The "missing" category due to independent censoring is handled through inverse probability weighting. Various competing risks models are considered, such as the Fine and Gray, multistate, and multinomial logistic models. Estimation methods for the NRI and the IDI from competing risks data are presented. The inference for the NRI is constructed based on asymptotic normality of its estimator, and the bias-corrected and accelerated bootstrap procedure is applied for the IDI inference. Simulations demonstrate that the proposed inferential procedures perform very well. The Multicenter AIDS Cohort Study is used to illustrate the practical utility of the extended NRI and IDI for competing risks outcomes.

EO484 Room MAL 153 COMMUNITY DETECTION, QUANTILE REGRESSION AND SURVIVAL ANALYSIS Chair: Oscar H. Madrid Padilla

E0678: **Community detection on social network with complex attributes**
Presenter: **Wanjie Wang**, National University of Singapore, Singapore

Social network analysis gains more and more interest in current studies, together with the case that nodes also have attributes. There are many studies on how to combine the network data and attribute data to achieve better community detection results, yet most of them are for low-dimensional attributes. In the statistician citation network data, we take the paper abstracts as the attribute data, which is high-dimensional and sparse. We proposed a new community detection method for such data with complex attributes, based on the SCORE method. We apply the method to the network data and get interesting results. We also prove corresponding theoretical results on the algorithm.

E1467: **High dimensional latent quantile regression**
Presenter: **Oscar H. Madrid Padilla**, UCLA, United States

A novel estimator is proposed for high dimensional latent quantile regression. The approach consists in combining the quantile loss function with recent advances in convex optimization. The resulting estimator can naturally be used for applications where there is interest in variable selection in high dimensions (perhaps p larger than n), but in the presence of low-rank latent factors. On the theoretical side, we consider a setting that allows for lagged dependency and show that, under suitable regularity conditions, for a fixed quantile level, our estimator can consistently estimate both, the vector of coefficients and the latent factor matrix. On the computationally side, we provide a solution algorithm based on the popular alternating method of multipliers. Finally, our experiments in real data show the value of our proposed method for interpretation of the latent factors, and variable selection.

E1479: **A quantile localized approach to the accelerated failure time model on survival data with time-dependent covariates**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong

The purpose is to discuss a generalization of the accelerated failure time model for survival data subject to right censoring, which is independent of the actual lifetime conditional on possibly time-varying covariates. We require the homogeneous conditional quantile assumption on the lifetime for a localized range of quantile levels, instead of assuming it hold globally. By introducing a class of weighted rank-based estimation procedure, the framework allows a quantile localized inference on the covariate effect with less stringent assumption. Meanwhile, the form of the proposed estimating equations can be viewed as a generalization of its counterpart under the accelerated failure time model with time-varying covariates. Numerical studies demonstrate that the proposed estimator overperforms current alternatives under various settings in terms of smaller empirical bias and standard deviation. A perturbation-based resampling method is also provided to reconcile the asymptotic distribution of the parameter estimates. Finally, consistency and weak convergence of the proposed estimator is established via empirical process theory.

E1986: **Adaptive community detection via fused l1 penalty**
Presenter: **Yunjin Choi**, University of Seoul, Korea, South

In recent years, community detection has been an active research area in various fields including machine learning and statistics. While a plethora of works has been published over the past few years, most of the existing methods depend on a predetermined number of communities. Given the

situation, determining the proper number of communities is directly related to the performance of these methods. Currently, there does not exist a golden rule for choosing the ideal number, and people usually rely on their background knowledge of the domain to make their choices. To address this issue, we propose a community detection method that also adaptively finds the number of the underlying communities. Central to our method is fused l-1 penalty applied on an induced graph from the given data. The proposed method shows promising results.

EO120 Room MAL 254 ASYMPTOTIC AND COMPUTATIONAL METHODS FOR STOCHASTIC PROCESSES Chair: Nakahiro Yoshida

E0462: **Rate of estimation for the stationary distribution of stochastic damping Hamiltonian systems with continuous observation**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Nakahiro Yoshida, Sylvain Delattre
The problem of the non-parametric estimation of the stationary measure π of a stochastic two dimensional damping Hamiltonian system $(Z_t)_{t\in[0,T]} = (X_t, Y_t)_{t\in[0,T]}$ is studied. From the observation of the path on $[0,T]$ we determine the rate of estimation of $\pi(x_0, y_0)$ as $T \to \infty$: we obtain a minimax lower bound on the estimation risk for pointwise estimation, and we show that this lower bound can be obtained by some estimators. One finding is that the rate of estimation is different with the one appearing in the standard i.i.d. setting or in the case of two dimensional non degenerate diffusion processes.

E0643: **Intensity ratios of marked point processes for limit order book modeling**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
A multidimensional point process with "Cox-type" intensities is considered. Given $\lambda_0(t)$ is an unobserved unspecified stochastic baseline intensity, and the X_j's are observable covariate processes, the intensity of the ith coordinate process is $\lambda^i(t) = \lambda_0(t)\exp\left(\sum_j \vartheta^i_j X_j(t)\right)$. In a previous work, we have proposed an estimation procedure of the parameters $\theta^i_j = \vartheta^i_j - \vartheta^0_j$ based on the quasi-likelihood of intensity ratios. Quasi-maximum likelihood estimators of θ^i_j's are consistent and asymptotically normal. This framework is suitable to model high-frequency order flows on a financial exchange. It provides a meaningful modeling of order submission intensities, an assessment of trading signals, and may have good prediction properties. We now extend the previous framework to the case of marked point processes. We are thus able to model arrivals of orders in a limit order book along with their size. We propose a multi-step ratio estimation procedure to sign market orders and determine whether they lead to a price change. The fitted model is able to provide out-of-sample predictions of the sign of the next price change (in a theoretical setting without any latency, computational cost or trading costs).

E0960: **Towards coding of the asymptotic expansion formula in YUIMA**
Presenter: **Emanuele Guidotti**, University of Neuchatel, Switzerland
Co-authors: Nakahiro Yoshida
Based on the Malliavin-Watanabe theory, a general asymptotic expansion formula designed has been previously presented to facilitate implementation. The aim is to show the current advances in the implementation of Asymptotic Expansion in YUIMA, open source academic project aimed at developing a complete environment for estimation and simulation of Stochastic Differential Equations and other Stochastic Processes via the R package called yuima and its Graphical User Interface yuimaGUI. The implementation of the asymptotic expansion formula would allow the user to efficiently compute expected values of multidimensional processes with virtually any degree of accuracy.

E1591: **Metropolis-within-piecewise deterministic Markov processes**
Presenter: **Kengo Kamatani**, Osaka University, Japan
Co-authors: Naohisa Okamoto
The Metropolis-Hastings (MH) jump update to the piecewise deterministic Markov processes (PDMP) is presented. PDMPs are useful for Bayesian computation, allowing unbiased subsampling. However, the implementation of PDMP requires efficient coding of the jump times. By combining the MH scheme, we can bypass this difficulty for some parameters of interest. We will show some theoretical properties of Markov processes defined by Metropolis-within-PDMPs. Finally, we apply it to Bayesian inference for stochastic processes.

EO641 Room Senate ADVANCES IN TEMPORAL EXTREMES Chair: Kirstin Strokorb

E0674: **A nonparametric estimator of the extremal index**
Presenter: **Juan Juan Cai**, Delft University of Technology, Netherlands
Co-authors: Andrea Krajina
Clustering of extremes usually has a large societal impact. The extremal index, a number in the unit interval, is a key parameter in modelling the clustering of extremes. We use a tool from multivariate extreme value theory to represent the extremal index, that is, we build a connection between the extremal index and the stable tail dependence function, which enables us to compute the value of extremal indices for some time series models. We also construct a nonparametric estimator of the extremal index using this connection. We prove that the estimator is consistent and asymptotically normal. The simulation study compares our estimator to the existing ones, which shows that our method has good finite sample properties. We also illustrate our method to a real data set on the daily maximum temperature in the Netherlands.

E1142: **Asymptotic analysis of subcritical branching processes with regularly varying immigration**
Presenter: **Bojan Basrak**, University of Zagreb, Croatia
When considering stationary multivariate regularly varying time series, it is useful to observe that, conditionally on the event that the norm of the present value exceeds a given threshold x, the whole sequence normalized by x has a limiting distribution as $x \to \infty$. That limit is called the tail process. Provided that time series satisfies some weak dependence conditions, its extremal behavior can be elegantly characterized using the notion of the tail process and the theory of point processes. However, except in a few simple cases, establishing such conditions and determining exact distribution of the tail process in the multivariate setting remains a technically challenging task. We study a class of models where we show a somewhat different route to asymptotic analysis. The motivation comes from the study of conditional least squares estimator of the mean number of progeny in the branching process with heavy tailed immigration. We also provide the rate of convergence and precise asymptotic distribution of the estimator.

E1228: **On trend estimation and testing with application to extreme rainfall**
Presenter: **Claudia Neves**, University of Reading, United Kingdom
Extreme Value Theory provides a rigorous mathematical justification for being able to extrapolate outside the range of the sampled observations. The primary assumption is that the observations are independent and identically distributed. Although the celebrated extreme value theorem still holds under several forms of weak dependence, relaxing the stationarity assumption, for example by considering a trend in extremes, leads to a changeling problem of inference based around the frequency of extreme events. Some studies have deemed that climate change is not so much about startling magnitudes of extreme phenomena, but rather how the frequency of extreme events can contribute to the worst case scenarios that could play out on the planet. For instance, the average rainfall may not be changing much, but heavy rainfall may become significantly more or less frequent, meaning that different observations must be endowed with different aspects in their underlying distributions. We will present statistical tools for semi-parametric modelling of the evolution of extreme values over time and/or space by considering a trend on the frequency of

high exceedances. The methodology is illustrated with an application to daily rainfall data from several gauging stations across Germany and The Netherlands.

E1336: **Hidden tail chains**
Presenter: **Ioannis Papastathopoulos**, University of Edinburgh, United Kingdom

Some key extremal features for kth order Markov chains are derived, which can be used to understand how the process moves to and fro between the body of the process and an extreme state. The chains are studied given that there is an exceedance of a threshold, as the threshold tends to the upper endpoint of the distribution. The extremal properties of the Markov chain at lags up to k are determined by the kernel of the chain, through a joint initialisation distribution, with the subsequent values determined by the conditional independence structure through a transition behaviour. We study the extremal properties of each of these elements under weak assumptions for broad classes of extremal dependence structures and show that it is possible to find a simple affine normalization, dependent on the threshold excess, such that non-degenerate limiting behaviour of the process is assured for all lags. These normalization functions have an interesting structure that has a striking parallel to the Yule-Walker equations. Furthermore, the limiting process is always linear in the innovations. We illustrate the results with the study of kth order stationary Markov chains based on widely studied families of k+1 dimensional copula.

EO374 Room MAL 251 Y-SIS: FROM METHODOLOGY TO APPLICATIONS Chair: Alessia Caponera

E1143: **Bayesian multiscale mixture of Gaussian kernels for density estimation**
Presenter: **Marco Stefanucci**, University of Padua, Italy
Co-authors: Antonio Canale

Some results related to a novel Bayesian nonparametric method for multiscale density estimation are discussed. Specifically, we extend a model originally developed for compact sample spaces to deal with data taking values in the whole real line R. By means of an infinitely-deep binary tree of kernels, we are able to construct a multiscale mixture model able to approximate densities with varying degrees of smoothness and local features. Sampling from the posterior distribution is available with a Markov Chain Monte Carlo method.

E0835: **Functional control charts based on scalar-on-function linear model**
Presenter: **Christian Capezza**, University of Naples Federico II, Italy
Co-authors: Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, Simone Vantini

Motivated by a real case study on monitoring ship operating conditions and CO2 emissions using navigation data from a roll-on/roll-off passenger cruise ship, functional control charts are proposed for monitoring multivariate profiles (i.e., multivariate functional data), which describe ship operating conditions at each voyage, and for prediction of the total CO2 emissions through scalar-on-function regression. The first contribution is the integration of functional regression into statistical process control techniques in a unified framework. Moreover, the proposed framework is suitably capable of making predictions and giving indications about possible anomalies in real-time, i.e., before the end of a voyage.

E0900: **Model based geostatistics and decision making: A parasite tale**
Presenter: **Claudio Fronterre**, Lancaster University, United Kingdom

Spatial analysis has increasingly become a valuable tool in the field of parasitic diseases. Particularly when disease prevalence is highly spatially heterogeneous, quantifying the spatial variability of disease risk and its uncertainty is crucial to inform disease control and elimination programmes. The complex framework that characterises several prevalence surveys often translates in a series of statistical challenges. Model based geostatistics can hugely help to address these issues. The applications shown will focus on the control and elimination of neglected tropical diseases (NTDs) affecting populations in sub-Saharan Africa.

E0875: **Spatial M-quantile regression with covariate measurement error to model housing price in Milan**
Presenter: **Francesco Schirripa Spagnolo**, Universita di Pisa, Italy
Co-authors: Riccardo Borgoni , Antonella Carcagni, Alessandra Michelangeli, Nicola Salvati

Spatial data have become increasingly common in order to study urban dynamics. However, these kinds of data are often affected by measurement error (ME) due to the bias induced during the data collection processes or for the statistical pre-processing often necessary to estimate the variables of interest at the desired spatial scale. If measurement error is ignored, standard regression estimation techniques may give biased regression coefficients. Moreover, the presence of outliers and influential points in the data can invalidate the assumptions of the classical regression models requiring the adoption of robust methods. A semiparametric M-quantile approach is proposed in order to obtain both bias-corrected and robust estimates of regression parameters. Moreover, this approach allows us to study the differential effect of a covariate at different levels of the conditional distribution of the response variable. The proposed methodology is applied to housing price in Milan, Italy. In particular, the main aim is to study the effect of cultural amenities and related infrastructures on the price levels.

EO679 Room MAL 252 NONPARAMETRIC STATISTICS: BAYESIAN AND FREQUENTISTS Chair: Mayer Alvo

E0399: **A new nonparametric tail risk measure**
Presenter: **Philip Yu**, The University of Hong Kong, Hong Kong
Co-authors: Keith Law, Wai-keung Li

The proposition of tail risk as a new asset pricing factor has gained traction in recent years. Recent work proxies the cross-sectional variation of returns by Fama-French portfolios and summarizes the cross-sectional variation by the PCA method to further reduce the dimension to a few basis assets. They then set the state-of-natures to be higher than the basis assets to estimate the stochastic discount factors for risk neutralizing the excess expected shortfall, which is taken as a tail risk measure. As an alternative approach to this double dimension reduction, we develop a nonparametric risk measure by directly forming portfolios which minimize the excess expected shortfall nonparametrically. Our empirical results reveal that the direct measure exhibits higher explanatory power when applied to more liquid and non-lottery style of stock returns.

E0731: **Modeling recurrent gap times and conditional estimating equations**
Presenter: **Ioana Schiopu-Kratina**, University of Ottawa, Canada
Co-authors: Hai Yan Liu, Mayer Alvo, Pierre-Jerome Bergeron

A semiparametric approach to the analysis of data from right censored recurrent events processes is presented. The dependence of the gap time between consecutive events on a set of covariates is explored. While the entire distribution of each gap time is not modeled, a regression-like dependence is specified for their conditional mean and variance. Under certain conditions on censoring, one can construct normalized estimating functions that are asymptotically unbiased and contain only observed data. Based on these estimating functions one can set up appropriate equations, which are a particular instance of generalized estimating equations. Solutions to these equations provide estimators for the regression and over dispersion parameters. Modern mathematical techniques are used to prove the existence, consistency and asymptotic normality of a sequence of such estimators. Examples that illustrate the theoretical approach are given. Numerical examples with simulated data are presented. A comparison of this methodological and technical approach to other comparable work in the field is presented. Conclusions based on the numerical results are also presented.

E1071: **Some new developments in nonparametric Bayesian inference**
Presenter: **Mahmoud Zarepour**, University of Ottawa, Canada

The Ferguson Dirichlet process introduces a prior on space of all probability measures. This prior is a random discrete probability measure which is dense over the space of all probability measures. The Dirichlet prior works like frequentists empirical process and this can be confirmed through asymptotic theory. For example, the Bayesian bootstrap shows asymptotic equivalence and an analogous *m* out of *n* Bayesian bootstrap can also be introduced which works like the frequentists regular *m* out of *n* Bayesian bootstrap. The goal is to present an overview of topics from nonparametric Bayesian inference which has an equivalent development in frequentist's paradigm. Some other important priors that work like Dirichlet process but with more flexibility will be introduced and a brief historical overview will be provided. Data augmentation and its applications to Machine Learning. We also provide some extensions to derive priors coming from many other general infinitely divisible processes (both uni-variate and multivariate).

E0213: **Estimating the local false discovery rate via a bootstrap solution to the reference class problem**
Presenter: **Mayer Alvo**, University of Ottawa, Canada
Co-authors: Farnoosh Abbas-Aghababazadeh, David R Bickel

Methods of estimating the local false discovery rate (LFDR) have been applied to different types of datasets such as high-throughput biological data, diffusion tensor imaging (DTI), and genome-wide association (GWA) studies. We present a model for LFDR estimation that incorporates a covariate into each test. Incorporating the covariates may improve the performance of testing procedures, because it contains additional information based on the biological context of the corresponding test. This method provides different estimates depending on a tuning parameter. We estimate the optimal value of that parameter by choosing the one that minimizes the estimated LFDR resulting from the bias and variance in a bootstrap approach. This estimation method is called an adaptive reference class (ARC) method. We consider the performance of ARC method under certain assumptions on the prior probability of each hypothesis test as a function of the covariate. We prove that, under these assumptions, the ARC method has a mean squared error asymptotically no greater than that of the other method where the entire set of hypotheses is used and assuming a large covariate effect. In addition, we conduct a simulation study to evaluate the performance of estimator associated with the ARC method for a finite number of hypotheses. We apply the proposed method to coronary artery disease (CAD) data taken from a GWA study and diffusion tensor imaging (DTI) data.

EO526 Room MAL 253 GRAPHICAL MODELS AND APPLICATIONS Chair: Kim-Anh Do

E0961: **A loss-based prior for Gaussian graphical models**
Presenter: **Laurentiu Catalin Hinoveanu**, University of Kent, United Kingdom
Co-authors: Fabrizio Leisen, Cristiano Villa

Gaussian graphical models have been used across various contexts to infer the conditional independence structure arising in the sampling distribution. In the Bayesian framework, the process of learning the structure is based on model selection, where the graph prior plays an important role. In the past, the discrete uniform distribution has usually been taken as the respective graph prior, but it suffers from assigning excessive mass on medium-sized graphs. Alternative priors have been proposed to alleviate this problem by considering model edge inclusions as independent Bernoulli variables, whilst also trying to foster sparse graphs. We will illustrate a graph prior based on a methodology involving loss functions which has the peculiarity of being tuned to represent a large palette of prior sparsity knowledge. We show the behaviour of the prior through simulation studies and real data analysis.

E0973: **Exploratory and confirmatory Bayesian tests for Gaussian graphical models**
Presenter: **Joris Mulder**, Tilburg University, Netherlands
Co-authors: Donald Williams, Luis Pericchi, Phillip Rast, Joris Mulder

Novel statistical methods are introduced for Bayesian hypothesis testing under Gaussian graphical models. These types of models assume a network structure between the outcome variables where an edge between two variables implies a nonzero partial correlation between the respective variables given the other variables. Exploratory Bayes factors are presented for testing whether a partial correlation is zero, positive, or negative. Confirmatory Bayes factors are presented to test specific network structures using equality and/or order constraints on the partial correlations. When the interest is in comparing network structures across multiple independent groups (e.g., a placebo group vs a treatment group) new posterior predictive checks and Bayes factor tests are presented. The Bayes factors are based on proper matrix F priors. The posterior predictive checks are based on the Kullback-Leibler divergence across different network models. The methodology is implemented in the R package BGGM and illustrated in an application on post-traumatic stress disorder.

E1557: **Bayesian learning of weakly structural Markov graph laws using sequential Monte Carlo methods**
Presenter: **Felix Rios**, The royal institute of technology, Sweden

A sequential sampling methodology for weakly structural Markov laws, arising naturally in a Bayesian structure learning context for decomposable graphical models, is presented. As a key component of the suggested approach, we show that the problem of graph estimation, which in general lacks natural sequential interpretation, can be recast into a sequential setting by proposing a recursive Feynman-Kac model that generates a flow of junction tree distributions over a space of increasing dimensions. We focus on particle McMC methods to provide samples on this space, in particular on particle Gibbs (PG), as it allows for generating McMC chains with global moves on an underlying space of decomposable graphs. The suggested sampling methodology is illustrated through numerical examples demonstrating high accuracy in Bayesian graph structure learning in both discrete and continuous graphical models.

E1181: **Graphical criteria for efficient total effect estimation in causal linear models**
Presenter: **Emilija Perkovic**, University of Washington, United States
Co-authors: Leonard Henckel, Marloes Maathuis

Covariate adjustment is commonly used for total causal effect estimation. In recent years, graphical criteria have been developed to identify all covariate sets that can be used for this purpose. Different valid adjustment sets typically provide causal effect estimates of varying accuracies. We introduce a graphical criterion to compare the asymptotic variance provided by certain valid adjustment sets in a causal linear model. We employ this result to develop two further graphical tools. First, we introduce a simple variance reducing pruning procedure for any given valid adjustment set. Second, we give a graphical characterization of a valid adjustment set that provides the optimal asymptotic variance among all valid adjustment sets. Our results depend only on the graphical structure and not on the specific error variances or the edge coefficients of the underlying causal linear model. They can be applied to DAGs, CPDAGs and maximally oriented PDAGs. Furthermore, the pruning procedure can be applied to MAGs and PAGs.

EO727 Room SH349 DYNAMIC TIME SERIES MODELLING Chair: Pramita Bagchi

E0993: **Bayesian spectral analysis of replicated time series**
Presenter: **Zeda Li**, City University of New York, United States

Technological advances have facilitated an explosion in the number of studies that collect time series data from multiple subjects to better understand how power spectra are associated with cross-sectional covariates. However, analyzing such data poses significant challenges due to the complicated structure of the power spectrum and the dynamic dependence structure between power spectra. While methods for single time series are rather extensive, existing methods for estimating the time-varying spectrum of a replicated time series are relatively few. We will introduce a

flexible spectral analysis framework for replicated time series and explore open research questions in replicated spectral analysis brought about by complicated modern data structures.

E1157: **A test for separability in covariance operators of random surfaces**
Presenter: **Pramita Bagchi**, George Mason University, United States

The assumption of separability is a simplifying and very popular assumption in the analysis of spatio-temporal or hypersurface data structures. It is often made in situations where the covariance structure cannot be easily estimated, for example because of a small sample size or because of computational storage problems. We propose a new and very simple test to validate this assumption. Our approach is based on a measure of separability which is zero in the case of separability and positive otherwise. We derive the asymptotic distribution of a corresponding estimate under the null hypothesis and the alternative and develop an asymptotic and a bootstrap test, which are very easy to implement. In particular, the approach does neither require projections on subspaces generated by the eigenfunctions of the covariance operator nor distributional assumptions as recently used by other works to construct tests for separability. We investigate the finite sample performance by means of a simulation study and also provide a comparison with the currently available methodology. Finally, the new procedure is illustrated analyzing a data example.

E1240: **Understanding brain network through linear latent variable models of functional connectivity**
Presenter: **Sandipan Roy**, University of Bath, United Kingdom

Neuroimaging-driven prediction of brain age, defined as the predicted biological age of a subject using only brain imaging data, is an exciting avenue of research. We seek to build models of brain age based on functional connectivity while prioritizing model interpretability and understanding. This way, the models serve to both provide accurate estimates of brain age as well as allow us to investigate changes in functional connectivity which occur during the ageing process. The methods proposed consist of a two-step procedure: first, linear latent variable models, such as PCA and its extensions, are employed to learn reproducible functional connectivity networks present across a cohort of subjects. The activations within each network are subsequently employed as features in linear regression model to predict brain age. The proposed method is employed on the data from the CamCAN repository and the inferred brain age models are further demonstrated to generalize using data from two open-access repositories: the Human Connectome Project and the ATR Wide-Age-Range.

E1335: **MACE: Multiscale abrupt change estimation under complex temporal dynamics**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Zhou Zhou

The focus is on the problem of detecting abrupt changes in trend whilst the covariance and higher-order structures of the system can experience both smooth and abrupt changes over time. The number of jump points is allowed to diverge to infinity with the jump sizes possibly shrinking to zero. The method is based on a multiscale application of an optimal jump-pass filter to the time series, where the scales are dense between admissible lower and upper bounds. The MACE method is shown to be able to detect all jump points within a nearly optimal range with a prescribed probability asymptotically. For a time series of length n, the computational complexity of MACE is $O(n)$ for each scale and $O(n\log^{1+\varepsilon}n)$ overall, where ε is an arbitrarily small positive constant. Simulations and data analysis show that, under complex temporal dynamics, MACE performs favourably compared with some of the state-of-the-art change point detection methods.

EC802 Room MAL 354 CONTRIBUTIONS IN TIME SERIES II Chair: Liudas Giraitis

E0499: **An integer-valued autoregressive process for seasonality**
Presenter: **Andrius Buteikis**, Vilnius University Faculty of Mathematics and Informatics, Lithuania
Co-authors: Remigijus Leipus

An integer-valued autoregressive process of order 1 for seasonality with period d and intra-seasonally dependent innovations ($\text{SINAR}(1)_d$) is proposed. Model properties are provided for the univariate and multivariate representation of the process. A computationally fast estimation method, which is based on conditional least squares with parameter restrictions, is proposed for the multivariate model representation and compared with a likelihood-based estimation method via the Monte Carlo simulation. An empirical application on different types of Chicago crime data is carried out in order to assess whether the proposed model is able to capture adequately the seasonality patterns in non-synthetic data.

E2004: **Forecast of the trend-cycle component in the frequency domain**
Presenter: **Consuelo Nava**, University of Aosta Valley, Italy
Co-authors: Maria Grazia Zoia

The frequency domain representation of a stochastic process clears the way to a filter-based approach to estimate and predict the trend-cycle component (TCC) of an economic time series. A novel methodology which hinges on a truncated ideal-low pass filter together with a stylized power spectrum (SPS) analyze, is introduced to forecast the TCC. The filter is obtained as a finite approximation of a double infinite Toepliz matrix with sinc functions as entries, to approximate the transfer function of the intended ideal filter. The SPS analyzer is meant to properly locate the cut-off frequency of the filter which must tally with the upper limit frequency of the bandwidth of the TCC. The latter depends on cycle features like evolutiveness, which may yield to a broadening of the original pertinent frequency band. The SPS analyzer, which splits the (average) power of a series throughout the frequency axis, allows to determine the contribution to the power (variability) of the TCC. It is shown how values of the TCC can be duly estimated and predicted by virtue of the almost idempotency of the selected filter and its characteristics. Under suitable assumptions on the erratic component of the series, confidence bounds for the TCC can be estimated in and out of the sample. An application to economic data shows the excellent performance of this approach, whose outcome compares favorably with those of other extant procedures.

E1701: **Real time prediction of irregular periodic time series data**
Presenter: **Chi Tim Ng**, Chonnam National University, Korea, South

By means of a novel time-dependent cumulated variation penalty function, a new class of real-time prediction methods is developed to improve the prediction accuracy of time series exhibiting irregular periodic patterns, in particular, the breathing motion data of the patients during the robotic radiation therapy. The proposed methods are designed so that real-time updates can be done efficiently with O(1) computational complexity upon the arrival of a new signal without scanning the old data repeatedly. The performances are tested via simulation under models involving abrupt changes and gradual changes in mean, trend, amplitude, and frequency.

E1746: **Uncertainty quantification for parameters and time series forecasting based on data assimilation**
Presenter: **Hiromichi Nagao**, The University of Tokyo, Japan
Co-authors: Shin-ichi Ito

Data assimilation (DA) is a computational technique that integrates numerical simulation models and observation data based on Bayesian statistics. DA is mainly applied in the weather forecasting, in which the simulation model consists of a set of differential equations that describe time evolution of the Earth's atmosphere. DA has been expanding its application fields to various areas such as seismology, biology and materials science. When DA is applied to a large-scale simulation model, the four-dimensional variational method (4DVar) is often used to optimize parameters and initial conditions in the simulation model, rapidly and accurately computing the derivative of a cost function that measures the difference between the model and data. The conventional 4DVar can obtain only the optima of the parameters and time series forecasting but never evaluates their uncertainties. We propose a new 4DVar that enables us to evaluate the uncertainties by using the second-order adjoint method. We demonstrate the validity of the proposed method by applying to the phase-field models, which are Allen-Cahn-type differential equations often used to simulate the

time evolution of grain growths in materials.

CO789 Room MAL B02 QUANTITATIVE ASSET MANAGEMENT Chair: Gaelle Le Fol

C0675: **Diversifying trends**
Presenter: **Charles Chevalier**, Universite Paris Dauphine, France
Co-authors: Serge Darolles

A new method is provided to disentangle the systematic component from the idiosyncratic part of the risk. A semi-parametric approach, associated with standard dimension reduction techniques, enables us to extract the common trending part of any financial asset. We apply this methodology on a large set of futures, covering all the major asset classes, and we extract a common Break risk factor. We first show that interrelations as we model them are higher for some cross-asset class combinations than from intra-asset class ones, such as JPYUSD and Gold. This result can be helpful when creating sectors in a portfolio diversification context, especially for dynamic strategies. Moreover, we prove the economic value of our contribution by applying a simple time-series momentum strategy on the extracted risk factor. The significant alpha exhibited is a proof of the risk premium, which we interpret as the optimal way to harvest macroeconomic trends.

C0773: **Investor sentiment and intraday bitcoin returns**
Presenter: **Thomas Renault**, Universita Paris 1 Panthaon-Sorbonne, France
Co-authors: Dominique Guegan

The purpose is to use a dataset of several million messages sent on Twitter and on Stocktwits to explore the relation between investor sentiment on social media and intraday Bitcoin returns. Computing returns and investor sentiment at various frequencies, from 1 minute to 24 hours, we do not find any strong evidence that lagged sentiment or lagged returns predict Bitcoin returns on recent periods (2016-2019). This result is robust to the method used to compute investor sentiment (lexicon-based approach and machine learning) and to the inclusion of news from the website CoinDesk.com. The Bitcoin market seems to be more efficient than is commonly understood.

C1183: **Evidence from a horse-race on the top of intra-daily forecasting models for algorithmic trading**
Presenter: **Beatrice Sagna**, Universite Paris Dauphine PSL, France

For every trader operating within stocks markets, a good prediction of intra-daily volume is a big concern. The better this prediction is, the better his (her) financial performances will be. In the literature, not much attention has been dedicated on predicting intra-daily volumes. The core is to propose a cartography of four main intra-daily forecasting models for volume existing in the literature based on four models. The contribution is to challenge these models in a horse race by replicating them considering two main criteria: accuracy and speed execution. The estimations provide evidence that models that take advantage of cross-sectional variations give faster estimations and more accurate predictions of intra-daily volumes than models that only consider time series variations. We discuss econometric insights and microstructure phenomenons to support such results.

C1464: **The earnings-announcement-day news puzzle**
Presenter: **Nicolas Moreno**, HEC Liege - University of Liege, Belgium
Co-authors: Marie Lambert

Glamour stocks are particularly responsive to new information falling on earnings announcements (EA). Using a sentiment measure to proxy for firm-specific news content, we show that glamour stocks are almost twice as sensitive to EA-news releases than value stocks. The EA-news effect is not reversed over the quarter following the announcement, suggesting a permanent impact generated by informed traders. Conversely, value stocks are more affected by subsequent news shocks post-EA. If investors realize on EA that value-stock news convey a disappointing amount of information, in the sense that implications for future earnings and returns are less obvious than for glamour stocks, then they will command a premium for bearing greater post-EA idiosyncratic risk. Consistent with this prediction, we find that low reaction to EA-news is priced: Stocks with the lowest sensitivity to EA-news earn a significant 17.99% annualized return premium and exhibit greater idiosyncratic risk levels post-EA.

CO771 Room MAL B04 BAYESIAN ECONOMETRICS Chair: Catherine Forbes

C0494: **Updating variational Bayes: Fast sequential posterior inference**
Presenter: **Nathaniel Tomasetti**, Monash Universeity, Australia
Co-authors: Catherine Forbes, Anastasios Panagiotelis

Variational Bayesian (VB) methods usually produce posterior inference in a time frame considerably smaller than traditional Markov Chain Monte Carlo approaches. Although the VB posterior is an approximation, it has been shown to produce good parameter estimates and predicted values when a rich class of approximating distributions are considered. We propose Updating VB (UVB), a recursive algorithm used to update a sequence of VB posterior approximations in an online setting, with the computation of each posterior update requiring only the data observed since the previous update. An extension to the proposed algorithm, named UVB-IS, allows the user to trade accuracy for a substantial increase in computational speed through the use of importance sampling. The two methods and their properties are detailed in two separate simulation studies. Two empirical illustrations of the proposed UVBmethods are provided, including one where a Dirichlet Process Mixture model is repeatedly updated in the context of predicting the future behaviour of vehicles on a stretch of the US Highway 101.

C0634: **Asset pricing using time-frequency dependent network centrality**
Presenter: **Michael Ellington**, University of Liverpool, United Kingdom
Co-authors: Jozef Barunik

A framework is provided where time-frequency dependent network centrality links to expected excess return of financial assets. Noting that investors trade on different horizons, it is essential to understand how the connectedness of a system influences risk premium over the short-, medium- and long-term. Viewing the market as being generated by a time-varying parameter VAR model implies that a shock to the j-th asset is time-frequency dependent. This creates a network of time-frequency connections among all assets in the market. We propose a new measure of time-frequency dependent network centrality and apply this to all stocks listed on the S&P500. Our findings indicate that our time-frequency dependent measures significantly price assets; particularly over the longer-term.

C1013: **Copula stochastic volatility in oil returns: Approximate Bayesian computation with volatility prediction**
Presenter: **Concepcion Ausin**, Universidad Carlos III de Madrid, Spain
Co-authors: Audrone Virbickaite, Pedro Galeano

Modeling the volatility of energy commodity returns has become a topic of increased interest in recent years, because of the important role it plays in today's economy. We propose a novel copula-based Stochastic Volatility model which allows for asymmetric volatility persistence. We employ ABC estimation technique that is appropriate for such highly-nonlinear model. We carry out two simulation studies and show that ABC is a comparable alternative to standard MCMC based methods. Finally, we present a real data application using WTI and Brent oil returns and show that the proposed asymmetric models outperform the symmetric ones in- and out-of-sample in terms of volatility prediction accuracy.

C0871: **A Bayesian transformation model for forecasting asset returns**
Presenter: **Gelly Mitrodima**, LSE, United Kingdom

Jointly modelling a finite collection of quantiles over time is considered under a Bayesian nonparametric framework. To address the challenges of formal Bayesian quantile inference, we propose a flexible Bayesian transformation model. This allows the likelihood and the quantile function to

be directly calculated, and define a novel stationary process which can be "centred" over a parametric model. The model is very general and its structure allows us to derive sufficient conditions for stationarity of some important sub-models. The application of the model to simulated and real data via Markov chain Monte Carlo (MCMC) methods shows that the model performs well for a range of data generating mechanisms.

CO651 Room MAL B20 ADVANCES IN TIME SERIES AND PANEL DATA ECONOMETRICS Chair: Indeewara Perera

C0551: **Estimation and inference for spatial models with heterogeneous coefficients: An application to U.S. house prices**
Presenter: **Natalia Bailey**, Monash University, Australia
Co-authors: M Hashem Pesaran, Michelle Aquaro

The focus is on the problem of identification, estimation and inference in the case of spatial panel data models with heterogeneous spatial lag coefficients, with and without (weakly) exogenous regressors, and subject to heteroskedastic errors. A quasi maximum likelihood (QML) estimation procedure is developed and the conditions for identification of spatial coefficients are derived. Regularity conditions are established for the QML estimators of individual spatial coefficients, as well as their means (the mean group estimators), to be consistent and asymptotically normal. Small sample properties of the proposed estimators are investigated by Monte Carlo simulations for Gaussian and non-Gaussian errors, and with spatial weight matrices of differing degrees of sparsity. The simulation results are in line with the paper's key theoretical findings even for panels with moderate time dimensions and irrespective of the number of cross section units. An empirical application to U.S. house price changes during the 1975-2014 period shows a significant degree of heterogeneity in spill-over effects over the 338 Metropolitan Statistical Areas considered.

C1125: **Inference on the change point in high dimensional time series models**
Presenter: **Abhishek Kaul**, Washington State University, United States
Co-authors: Venkata Jandhyala, Stergios Fotopoulos

We develop a projected least squares estimator for the change point parameter in a high dimensional time series model with a potential change point. Importantly we work under the setup where the jump size may be near the boundary of the region of detectability. The proposed methodology yields an optimal rate of convergence despite high dimensionality of the assumed model and a potentially diminishing jump size. The limiting distribution of this estimate is derived, thereby allowing construction of a confidence interval for the location of the change point. A secondary near optimal estimate is proposed which is required for the implementation of the optimal projected least squares estimate. The prestep estimation procedure is designed to also agnostically detect the case where no change point exists, thereby removing the need to pretest for the existence of a change point for the implementation of the inference methodology. Our results are presented under a general positive definite spatial dependence setup, assuming no special structure on this dependence. The proposed methodology is designed to be highly scalable, and applicable to very large data. Theoretical results regarding detection and estimation consistency and the limiting distribution are numerically supported via monte carlo simulations.

C1271: **Bootstrap methods for multiplicative error models**
Presenter: **Indeewara Perera**, University of Sheffield, United Kingdom
Co-authors: Mervyn Silvapulle

The recent literature on time series analysis has devoted considerable attention to nonnegative time series, such as financial durations, realized volatility, and squared returns. The class of models, referred to as the Multiplicative Error Models (MEM), is particularly suited to model such nonnegative time series. A novel bootstrap-based method is proposed for producing multi-step-ahead probability forecasts for MEMs, including distributional forecasts. In order to test the adequacy of the underlying MEM, a class of bootstrap specification tests are also proposed. The proposed bootstrap methods are shown to be asymptotically valid. Monte Carlo simulations suggest that our methods perform well in finite samples. A real data example illustrates the methods.

C1308: **Bootstrap tests for integer-valued GARCH models with covariates**
Presenter: **Adriana Cornea-Madeira**, University of York, United Kingdom
Co-authors: Andreea Halunga

Estimation and testing on the boundary of the parameter space is nontrivial, particularly if the variable of interest is discrete. We derive the asymptotic distribution of the QMLE and Wald test for the integer-valued GARCH model with covariates, INGARCH-X, when parameters are on the boundary. The asymptotic distribution of the Wald test is non-standard and, in general, non-pivotal. As a consequence of this, coupled with the poor finite sample properties of the Wald test, we propose an asymptotically valid parametric bootstrap procedure for testing the statistical significance of parameters in an INGARCH-X model. The Monte Carlo study reveals that the bootstrap procedure performs really well. We apply these results to test for temporal clustering in extra-tropical cyclones.

CO777 Room MAL B35 FORECASTING WITH MANY PREDICTORS Chair: Daniel Borup

C0756: **Forecasting corporate earnings with machine learning**
Presenter: **Jorge Wolfgang Hansen**, Aarhus University and CREATES, Denmark
Co-authors: Christoffer Thimsen

A comparative empirical analysis is performed for a set of machine learning methods for predicting future firm-level earnings. In addition, we show how to incorporate monthly and quarterly data into the forecast models. We consider generalized linear models, dimension reduction, boosted regression trees, random forests, and neural networks. We find that the best performing machine learning models are superior to existing accounting-based predictive models and our findings are consistent across industries and firm characteristics. Utilizing intra-year data enhances also the practicability of our forecasts and makes them a useful alternative to the commonly applied analyst forecast. In particular, we show that the forecasts from the machine learning models are unbiased and outperform the usually over-optimistic financial analysts forecasts. Our findings show that machine learning significantly increases our ability to predict future economic outcomes and gives insight into where machine learning is particularly useful.

C1641: **A machine learning approach to volatility forecasting**
Presenter: **Mathias Siggaard**, Aarhus University, Denmark

This paper shows how machine learning algorithms can improve the forecast accuracy of one-day-ahead forecast for high-frequency for volatility series. To achieve this, all stocks from Dow Jones Industrial Average index over the sample period from 2001 to 2018 are examined. Four groups of machine learning algorithms are compared; Regularization Methods, Tree-Based Methods, Deep Learning, and Ensemble Methods. Comparison with the commonly-used Heterogeneous Autoregressive model shows substantial improvement. Through model confidence set and forecast comparison, the best machine learning algorithms (neural networks and random forest) are identified. Furthermore, it is shown how these methods are capable of extracting important information when including additional explanatory variables and find a small set of dominating predictors including implied volatility, earnings announcements, and a daily policy index for United States. To better understand how these methods behave under extreme situations, this paper investigates their performance during the flash crash of 24th of August 2015.

C0968: **The non-linear relationship between macro-financial state variables and market volatility**
Presenter: **Johan Jakobsen**, Nordea, Denmark
Co-authors: Daniel Borup, Bezirgen Veliyev

The relationship between the level of financial market volatility and macro-financial state variables is investigated. To this end, we propose a parametric alternative to the GARCH model, labelled the ST-GARCH model, in which the level (intercept) is allowed to be time-varying as a smooth function of those state variables. This model entertains non-linear relationship(s) between the state variable and volatility. We establish the asymptotic theory of the quasi-maximum-likelihood estimator and develop a testing framework for selecting the number of transitions. Our broad application to the equity, foreign exchange, fixed income, credit, and real estate markets document clear level shifts in volatility. As such, our ST-GARCH realizes significant in-sample gains by capturing those (non-linear) level shifts using macro-financial state variables. This also translates to notable out-of-sample gains.

C0771: **Forecasting house prices using online search activity**
Presenter: **Erik Christian Montes Schutte**, Aarhus University, Denmark

It is shown that Google search activity is a strong out of sample predictor of future growth in U.S. house prices and that it strongly outperforms standard predictive models based on macroeconomic variables as well as autoregressive models. We extract the most important information from a large set of search terms related to different phases of the home search process into a single Google based factor and then use it to predict movements in future house prices. At the one-month forecast horizon, the Google factor delivers an out of sample R2 statistic of about 50 percent for the aggregate U.S. market over the period 2009-2018. We show that the strong predictive power of Google search activity holds for longer forecast horizons, for various house price indices, for seasonally unadjusted and adjusted data, and across individual U.S. states.

CO252 Room G4 MODELING REGIME CHANGE II Chair: Willi Semmler

C1426: **Oil prices and banking instability: A jump-diffusion model for bank capital structure**
Presenter: **Samar Issa**, Saint Peters University, United States
Co-authors: Willi Semmler

An empirical model of bank capital structure is developed to study the impact of large oil shocks on overleveraging of banks which then present severe challenges for banks balance sheet management. The measure of overleveraging incorporates a jump-diffusion component that captures the jump size and intensity of oil prices and political instability predictors. Overleveraging is derived and estimated for a sample of six banks in three oil-producing countries and Western countries using Markov Chain Monte Carlo method, for the years 2006-2016. The estimation of the optimal debt shows that most of the banks in this context had a high optimal debt around 2008, overlapping with the oil price shock. In addition, most of the predictors, namely oil prices and political instability factors proxied by terrorism, political corruption, and military expenses, regularly appeared in volatility and jump intensity factors.

C1094: **Multivariate nonlinear analysis of quarterly China's GDP and world oil price**
Presenter: **Fredj Jawadi**, University of Lille, France

The relationship between quarterly China's GDP and world oil price over the recent decades is studied in order to investigate whether the growing Chinese economy has any significant influence on the world oil market and to measure to what extent oil shocks affect China's economy. To this end, we apply a multivariate threshold autoregressive model for the two time series and study the associated nonlinear impulse response functions. We also contrast the implications of nonlinear model with those of linear models. Our findings show certain strong nonlinear effects between China's GDP and the oil sector. In particular, our nonlinear VAR model points to the presence of nonlinear causality effects when the oil price was declining, suggesting further evidence of bilateral interactions.

C1783: **Non-stationary DSGE models with time-varying steady state**
Presenter: **Viktors Ajevskis**, Bank of Latvia, Latvia

DSGE models are designed to explain cyclical features of the data. There are two approaches to deal with data for estimating the parameters of DSGE models: a. filter the data using statistical filter and then estimate the structural parameters with the output of the filter; b. transform the data using model-based specification of what the non-cyclical component is, then estimate the structural parameters with the transformed data. In both cases it is assumed that there exists a fixed steady state (either in the level or in the growth rate) in the model and the Blanchard-Kahn conditions hold. However, both approaches have their own problems. In the proposed approach a DSGE model with a unit root process in technology is considered. In this case the fixed steady state solution does not exist. A time-varying steady state is defined as a solution to which the economy converges in the absence of the future shocks. Presenting the data as a sum of the time-varying steady state and deviation from that allows using the Kalman filter for estimating the model parameters and unobservable variables. A prototypical DSGE model and the US data to estimate the model parameters and the output gap is used.

C2022: **The US post-war economic dynamics and price/wage setting processes: A regime-switching DSGE approach**
Presenter: **Giovanni Di Bartolomeo**, Sapienza University of Rome, Italy
Co-authors: Elton Beqiraj

This paper reconsiders the evolution of the US business cycle by considering a Markow-switching (MS) approach. To capture the traditional debate between "good luck" and "good policy," we introduce the possibility of a regime switches in monetary policy parameters and in the variability of shocks. The novelty of our paper is also considering switches in the Phillips curves of wages and prices; in particular, we consider switches in the intrinsic inertia of wage and price inflation. The inertia is formalized through time-dependent price adjustments based on hazard functions that are not flat. From this point of view, our paper attempt to conciliate the literature on time-dependent to that of the state-dependent price adjustment. Our preliminary results can be summarized as follows. 1) Price and wage inflation intrinsic persistence matters for MS models (i.e., MS switches in the price and wage hazard). 2) Great Moderation is mainly explained by changes in the shock volatilities (good luck), once we control for intrinsic persistence of prices, MS monetary regimes do not account (good policies) for the observed macroeconomic changes, 3) Great Inflation is driven by changes in the price hazard. Price and wage setting institutions matter. 4) New interpretation of switches in monetary policy regime based on Brainard Principle and deviations from it based on current judgments.

CO859 Room Chancellor's Hall STATE-SPACE REPRESENTATIONS: COMPUTATION AND APPLICATIONS Chair: Frederic Karame

C1750: **Kalman filter: A Julia implementation**
Presenter: **Michel Juillard**, Bank of France, France

The Kalman filter and smoother is central to the estimation of linear state space models and of linearized DSGE models in particular. Estimation techniques, based either on optimization or on MCMC, require a large number of evaluations of the Kalman filer. This is the most time consuming single step in practical DSGE modeling. It is therefore essential to have a fast implementation of this algorithm. The Julia language is known for its speed and provides various tools to optimize code. It is a natural testing bench. Two variants of the filter are also explored: the Chandrasekhar recursion and the diffuse filter. An alternative implementation with PaddedMatrices is also presented.

C1790: **Online estimation of DSGE models**
Presenter: **Marco Del Negro**, Federal Reserve Bank of New York, United States

The usefulness of sequential Monte Carlo (SMC) methods in approximating DSGE model posterior distributions is illustrated. We show how the tempering schedule can be chosen adaptively, explore the benefits of an SMC variant we call generalized tempering for online estimation, and provide examples of multimodal posteriors that are well captured by SMC methods. We then use the online estimation of the DSGE model to

compute pseudo-out-of- sample density forecasts of DSGE models with and without financial frictions and document the benefits of conditioning DSGE model forecasts on nowcasts of macroeconomic variables and interest rate expectations. We also study whether the predictive ability of DSGE models changes when we use priors that are substantially looser than those commonly adopted in the literature.

C1785: **One-step online maximum likelihood for linear state-space representations**
Presenter: **Frederic Karame**, Le Mans University, France
Co-authors: Alexandre Brouste

State-space models can be difficult to estimate by maximum likelihood due to the usual numerical problems (size, slow calculations, local solutions, ...). The one-step online approach has the double advantage of circumventing the usual numerical problems and providing efficient estimators. Nevertheless, these properties have been obtained for rather simple models and not for complex models like linear state-space representations. The aim is to extend this online estimation method to linear state-space representations. An efficient and fast estimation of these models represents an important breakthrough, especially if it can be implemented transparently for a user. The first part of the paper presents the theoretical proof. The second part is devoted to some Monte Carlo experiments. In the third part, the method is implemented to macroeconomic or financial issues.

CO380 Room Jessel CHANGE POINT PROBLEMS IN STOCHASTIC PROCESSES: THEORY AND APPLICATIONS Chair: Ilia Negri

C1130: **Self-weighted GEL method for heavy-tailed ARMA models and its applications to various problems**
Presenter: **Fumiya Akashi**, The University of Tokyo, Japan

A testing problem of linear hypothesis on the coefficients of heavy-tailed ARMA processes is considered. It is well known that heavy-tail phenomena of time series models are known to cause some problem, such as intractable form of the rate of convergence or complicated limit distribution. As a result, it is often difficult to detect critical values of tests or cut-off points of confidence interval based on the limit distribution of test statistics in infinite variance cases. To overcome the difficulties, the aim is to construct the least absolute deviations regression and self-weighting-based generalized empirical likelihood (GEL) statistics for the testing problem of ARMA models. By the self-weighting and GEL, the proposed test statistic is shown to have a pivotal chi-squared limit distribution regardless of whether the model has infinite variance or not. In the latter half of this talk, we apply the self-weighted GEL method to the test of causality and change-point detection problem of heavy-tailed time series models.

C1425: **A change detection procedure for an ergodic diffusion process**
Presenter: **Koji Tsukuda**, The University of Tokyo, Japan

A test procedure to detect a change in values of drift parameters of an ergodic diffusion process is considered under the setting of continuous observation. For this problem, there is an approach based on the weak convergence of a random process relating to an estimating equation. Asymptotic null distributions of some test statistics are established by using this weak convergence result and the continuous mapping theorem. We show the weak convergence of a weighted version of the random process and propose a weighted test statistic.

C1285: **On smooth change-point estimation for Poisson processes**
Presenter: **Serguei Dachian**, Universite de Lille, France
Co-authors: Arij Amiri

Suppose n independent realizations of an inhomogeneous Poisson process are observed whose intensity function goes from one given level to another in a quick (but smooth) manner in the vicinity of an unknown point θ. The size δ_n of the vicinity is supposed to converge to zero as n goes to infinity. It turns out that the behavior of the maximum likelihood and Bayesian estimators (MLE and BEs) of θ strongly depends on the rate of convergence of δ_n to zero. We show that if this convergence is slow, the problem remains regular and the MLE and BEs are asymptotically normal (with a rate comprised between $1/\sqrt{n}$ and $1/n$) and asymptotically efficient. While if the convergence is fast, we show that — like in the discontinuous case — the rate of convergence of the estimators is $1/n$, their limiting laws are no longer Gaussian, and only BEs are asymptotically efficient.

C1211: **Change point detection based on method of moment estimators**
Presenter: **Ilia Negri**, University of Bergamo, Italy

A change point detection procedure using the method of moment estimators is proposed. The test statistics is based on a suitable Z-process. The asymptotic behavior of this process is established under both the null and the alternative hypothesis and the consistency of the test is also proved. An estimator for the change point is proposed and its consistency is derived. Some examples of this method applied to some parametric families of random variables are presented.

CC830 Room MAL 351 CONTRIBUTIONS IN EMPIRICAL FINANCE Chair: Cesare Robotti

C0710: **Financial markets' reactions to the Greek crisis' policy response: A behavioural finance approach**
Presenter: **Simon Ganem**, LSE, France

The purpose is to investigate the reaction of sovereign Greek spreads to the policy response to the Greek crisis between January 2010 and Mario Draghi's July 2012 "whatever it takes" speech. In particular, we try to determine to what extent financial markets behaved efficiently during those crisis times. In order to do so, we have constructed a dataset of thousands of manually coded news items. Manually coding news has allowed us to differentiate between expected and non-expected news as well as minimise the endogeneity bias that can undermine the reliability of existing econometric results. The analysis of daily data, through different econometric approaches, from standard OLS to event study or quantile regressions, would suggest that financial markets did not (entirely) react in an efficient manner to policy developments. Overall, markets were more sensitive to negative than positive news when all types of news were considered (decisions, rumours, media reports, statements). When looking at a more disaggregated level of analysis, our findings show that markets reacted only to unexpected decisions when positive in line with efficiency predictions. In the meantime, rumours, media reports, statements as well as expected decisions had an impact when negative, more in line with behavioural predictions.

C1791: **When the (expected) loss quantiles go marching in**
Presenter: **Maria Magdalena Vich Llompart**, Washington College, United States

Option-implied information has been proved to be more accurate in predicting future volatility, returns and downturns. Moreover, the information embedded in the tails is linked to macroeconomic variables. We analyze for the first time the international connectedness of option-implied loss quantiles, which are related with the price of the Arrow-Debreu assets in the worst scenarios. Using data from the S&P 500, the EuroStoxx 50, the Nikkei 225 and the FTSE 100 index options, we extract the risk-neutral densities and calculate different connectedness measures to confirm that US is the main transmitter of shocks in the S&P 500 risk-neutral loss quantile while Japan is the main receiver.

C1840: **Allocation choice between pension fund and long term care: The investor's perspective in the Italian market**
Presenter: **Nancy Zambon**, University of Padova, Italy
Co-authors: Massimiliano Caporin

Focusing on the Italian market, we investigate the optimal allocation choice of an investor who can choose to invest exclusively in pensions funds against the decision of reducing the contribution to the pension fund to underwrite a Long Term Care insurance. We first determine the optimal allocation in the pension fund accounting for uncertainty in the model parameters. Indeed, although the investment horizon is relevant in allocation choices, considering uncertainty in the parameters the horizon effect is less present. We then use this information to determine the cash flows

associated with different possible life events and investment strategies. This in turn is used as a measure of Long Term Care utility in terms of expected economic wealth differential between the cases with and without insurance. The results suggest that the rational choice for an Italian investor is not to subscribe a Long Term Care coverage. Indeed, when investors optimally allocate their wealth in the pension fund, accounting also for uncertainty, the expected advantage from the Long Term Care is negative.

C1906: **Predicting intraday return patterns based on overnight returns for the US stock market**

Presenter: **Hao Li**, University of Amsterdam, Netherlands

Co-authors: Cees Diks, Valentyn Panchenko

A new approach, cumulative regression (CumRe), is proposed in order to predict intraday financial return patterns conditional on observed overnight returns. Based on Trade and Quote data, we find evidence for dependence between overnight returns and subsequent intraday first and last half-hour return patterns for the S&P 500 Exchange-Traded Fund for the time period from 2003 to 2013 with both statistical and economic significance. Our methodology allows studying the return patterns documented in the existing theoretical and empirical literature in more detail. Moreover, we find that both the first and the last half hours offer opportunities for day traders. Specifically, 20-minute after the market opens, and 30-minute before the closing are the best times for trading in terms of annualized returns, Sharpe ratios, and the difference between the Certainty Equivalent Returns.

CC829 Room MAL 352 CONTRIBUTIONS IN ECONOMETRICS MODELLING Chair: Tso-Jung Yen

C1812: **Sovereign default of European countries: Evidence from truncated factor vine copula model of CDS Spreads**

Presenter: **Hoang Nguyen**, Orebro University, Sweden

Co-authors: Concepcion Ausin, Pedro Galeano

A truncated factor vine copula model is proposed to capture the joint sovereign default of European countries during and after the debt crisis. The truncated factor vine copulas can be considered as a combination of a factor copula model at the first tree layer and truncated vine copulas at higher tree layers. We employ the variational Bayesian approach to estimate copula parameters and incorporate a procedure to select the best bivariate links among the nodes of the model. We find that the dependence structure of sovereign default is fat tail and asymmetric. The conditional default probability of European countries fails if another European country fails peaks during the crisis. Furthermore, European countries become less integrated after the crisis.

C1837: **A Bayesian analysis of the extended Poisson regression**

Presenter: **Haruhiko Shimizu**, Kobe University, Japan

The Bayesian analysis of the extended Poisson regression model is considered. The support of the extended Poisson distribution includes not only the non-negative integers but also negative integers. The usual Poisson distribution is a special case of the extended Poisson distribution. We study the Bayesian estimation of the parameters of extended Poisson distribution and show that Bayesian analysis performed better than maximum likelihood method. In particular, Bayesian method was better in estimating the parameter of the ratio of positive or negative values in the dataset when the ratio is close to zero or one. In these cases, maximum likelihood method was not available since the Hessian did not converge for some cases. As for Bayesian method, we are able to estimate the ratio parameter if we choose the appropriate prior density. Based on the previous study, we next construct the extended Poisson regression model. We then compare the Bayesian method and maximum likelihood method for estimating the parameters of the extended Poisson regression model. We also consider an application of this model.

C1980: **The disposition effect: Behavioural evidences from a quantile regression with fixed effects and attrition**

Presenter: **Malvina Marchese**, Cass Business School, United Kingdom

Co-authors: Richard Payne

The disposition effect is investigated in an under-researched, but increasingly important field: retail currency traders. Thanks to the properties of the FX market and the availability of a large panel of traders, we model the disposition effect as the skewness of the conditional distribution of traders' returns. We introduce a novel fixed effect estimator for quantile regression in dynamic panel with fixed effects and attrition and establish its finite sample properties via an extensive simulation exercise. We find that the WPIVQRFE estimator significantly reduces the bias arising from the incidental parameter problem. We propose a jack-knife type correction in large samples to obtain asymptotic normality. The results indicate that realizing a loss on one trading day significantly increases the disposition effect on the next trading day, consistent with investors becoming even less likely to close a loss-making position. This is strong evidence that the disposition effect cannot be adequately explained by pure preference-based theories, and must instead be attributed to behavioural effects.

C1724: **Bayesian analysis of lognormal mixtures with an unknown number of components from grouped data**

Presenter: **Kazuhiko Kakamu**, Kobe University, Japan

A reversible jump Markov chain Monte Carlo method is proposed for estimating lognormal mixtures from grouped data. Using the posteriors we also consider the calculation of the Gini coefficient. Using both simulated and real data examples, we examined the performance of the proposed algorithm and the accuracy of the Gini coefficients. From the simulated data, we can confirm that the distributions are estimated accurately even in the case of grouped data and the Gini coefficients are also estimated very well. From the empirical example, we can identify that there exist two subgroups in 2012 in Japan.

Monday 16.12.2019 16:35 - 17:50 Parallel Session P – CFE-CMStatistics

EO560 Room CLO B01 FUNCTIONAL ANALYSIS FOR MULTIPLE TYPE DATA Chair: Xiongtao Dai

E1007: **Kernel regression with convolved Gaussian processes on Riemannina manifold**
Presenter: **Jinzhao Liu**, Newcastle University, United Kingdom

The data analysis on vector spaces is well studied. However, for some new and popular topic, such as computer vision and medical image analysis, the data are often mapped onto a special non-Euclidean space. Most of current models do not work since the data lack of vector structure. With the motivation of solve this problem, we try to find a model which can find the relationship between real-valued covariate and manifold-valued response variable. This model includes two part: one is the mean structure which is a generalisation of kernel method to Riemannian manifold, the other is covariance structure which is based on wrapped Gaussian process on Riemannian manifold. The purpose is to derive a concurrent model for manifold-valued data. In addition, uncertainty and random error are also considered in this mode.l

E1024: **Minimax powerful functional analysis of covariance tests for longitudinal genome-wide association studies**
Presenter: **Sheng Xu**, The Hong Kong Polytechnic University, China
Co-authors: Yehua Li, Catherine Liu

The Alzheimers Disease (AD) related phenotype response variables observed on irregular time points in longitudinal Genome-Wide Association Studies (GWAS) are modeled as sparse functional data and propose nonparametric test procedures to detect functional genotype effects while controlling the confounding effects of environmental covariates. Existing nonparametric tests do not take into account within-subject correlations, suffer from low statistical power and fail to reach the GWAS significance level. We propose a new class of functional analysis of covariance (fANCOVA) tests based on a seemingly unrelated kernel smoother, that can incorporate the correlations. We show that the proposed test combined with a uniformly consistent nonparametric covariance function estimator enjoys the Wilks property and is minimax most powerful. In an application to the Alzheimers Disease Neuroimaging Initiative data, the proposed test leads to the discovery of new genes that may be related to AD.

E1168: **Combination of multiple functional markers to improve diagnostic accuracy**
Presenter: **Qinyi Zhang**, The Hong Kong Polytechnique University, Hong Kong

Powerful diagnostic marker is important in diagnosis. Because of development of modern technique, diagnostic markers can be observed repeatedly and act as functional markers. Existing methods mainly discussed diagnosis by a single scalar marker or combinations of multiple scalar markers but not functional markers. Methods of functional data analysis are used to make diagnosis for functional markers. In particular, we adopt functional principal components analysis to obtain basis functions and the corresponding projections, derive the features on the basis of the projections, and finally the combinations of the obtained features. Receiver operating characteristic (ROC) curve is widely used for evaluating diagnosis, which can be assessed by area under the curve (AUC) or Youden Index. Our proposed methods are illustrated by simulations and real data analysis of diagnosis for high- or low- hospital admissions due to respiratory diseases in Hong Kong.

EO306 Room MAL B02 IMAGING GENETICS Chair: Mark Fiecas

E0232: **Bayesian GWAS with structured and non-local priors**
Presenter: **Adam Kaplan**, University of Minnesota Twin Cities, United States
Co-authors: Mark Fiecas, Eric Lock

A novel Bayesian approach to genome-wide association studies (GWAS) is introduced which improves over existing methods in two important ways. First, we describe a model that allows for a marker's gene-parent membership and other characteristics to influence its probability of association with an outcome. For this we use a hierarchical Dirichlet Process (DP) model that allows for clustering of the genes in tandem with a regression model for marker-level covariates. Second, we use Non-Local priors to model the difference in probability of minor allele status between patient disease status. We outline the implementation of and discuss the philosophical problems treated by Non-Local priors within the genome-wide analysis framework. In Bayesian hypothesis testing, it is often overlooked that the null hypothesis is a sub-event of the alternative hypothesis. This results in the asymptotic rates of convergence favoring the alternative hypothesis over the null, whereas we define a Non-Local prior for the GWAS context that gives symmetric rates of convergence. We assess the structured and Non-Local components with simulation studies under various scenarios. We apply our Bayesian GWAS method to single-nucleotide polymorphisms (SNP) data collected from a pool of Alzheimer's disease and cognitively normal patients from the Alzheimer's Database Neuroimaging Initiative.

E1216: **A simple, consistent estimator of SNP heritability from genome-wide association studies**
Presenter: **Armin Schwartzman**, University of California, San Diego, United States
Co-authors: Andrew Schork, Rong Zablocki, Wesley Thompson

Analysis of genome-wide association studies (GWAS) is characterized by a large number of univariate regressions where a quantitative trait is regressed on hundreds of thousands to millions of single-nucleotide polymorphism (SNP) allele counts, one at a time. An estimator is proposed for the SNP heritability of the trait, defined here as the fraction of the variance of the trait explained by the SNPs in the study. The proposed GWAS heritability (GWASH) estimator is easy to compute, highly interpretable, and is consistent as the number of SNPs and the sample size increase. More importantly, it can be computed from summary statistics typically reported in GWAS, not requiring access to the original data. The estimator takes full account of the linkage disequilibrium (LD) or correlation between the SNPs in the study through moments of the LD matrix, estimable from auxiliary datasets. Unlike other proposed estimators in the literature, we establish the theoretical properties of the GWASH estimator and obtain analytical estimates of the precision, allowing for power and sample size calculations for SNP heritability estimates, and forming a firm foundation for future methodological development.

E1555: **Covariate-modulated shrinkage estimator for imaging and genetics**
Presenter: **Wesley Thompson**, Institute of Biological Psychiatry, Denmark

Estimates of Total Variance Explained, termed "Heritability" in Genetics applications, are useful in many contexts. In scenarios where the amount of variance explained by a set of predictors is substantial, but effects of interest are broadly distributed across a high-dimensional explanatory variables, it becomes useful to determine attributes or features of the variables that are "enriched" for strength of association. Features themselves may be high-dimensional, including multiple discrete and/or continuous variables. Moreover, predictors (e.g., SNPs or vertices) may be correlated (e.g., due to linkage disequilibrium or to spacial smoothness). We describe a novel Bayesian global-local shrinkage algorithm to estimate overall variance explained in this setting, which incorporates potentially high-dimensional covariates. The effects of the covariates are themselves regularized. The model is fitted via an MCMC algorithm. We apply this methodology to genetic and imaging data from the Adolescent Brain and Cognitive Development (ABCD) study, a 12,000 strong population study of US children aged 9-10 years at baseline, examining the association of whole-genome genotyping and brain imaging data on neurocognitive outcomes.

EO530 Room MAL B04 RECENT ADVANCES IN METHODS FOR DYNAMIC TREATMENT REGIMES Chair: Kristin Linn

E0227: **A Bayesian imputation approach to optimizing dynamic treatment regimes**
Presenter: **Thomas Murray**, University of Minnesota, United States

Medical therapy often consists of multiple stages, with a treatment chosen by the physician at each stage based on the patient's history of treatments

and clinical outcomes. These decisions can be formalized as a dynamic treatment regime. This talk describes a new approach for optimizing dynamic treatment regimes that bridges the gap between Bayesian inference and Q-learning. The proposed approach fits a series of Bayesian regression models, one for each stage, in reverse sequential order. Each model uses as a response variable the remaining payoff assuming optimal actions are taken at subsequent stages, and as covariates the current history and relevant actions at that stage. The key difficulty is that the optimal decision rules at subsequent stages are unknown, and even if these optimal decision rules were known the payoff under the subsequent optimal action(s) may be counterfactual. However, posterior distributions can be derived from the previously fitted regression models for the optimal decision rules and the counterfactual payoffs under a particular set of rules. The proposed approach uses imputation to average over these posterior distributions when fitting each regression model. An efficient sampling algorithm, called the backwards induction Gibbs (BIG) sampler, for estimation is presented, along with simulation study results that compare implementations of the proposed approach with Q-learning.

E0725: **User-friendly estimation of optimal adaptive treatment strategies**
Presenter: **Erica Moodie**, McGill University, Canada

The goal of precision medicine is to tailor treatment strategies on an individual patient level. Although many estimation techniques are available, some are not robust to model mis-specification while others requite a steep learning curve to implement. A previous dynamic weighted ordinary least squares regression model was approachable yet robust, but was limited to binary exposures and continuous outcomes. We will discuss extensions to this approach that allow for multi-valued or continuous treatments, or for censored outcomes and demonstrate their performance in real and simulated data.

E1722: **Sample size considerations for comparing dynamic treatment regimes in a SMART with a longitudinal outcome**
Presenter: **Nicholas Seewald**, University of Michigan, United States
Co-authors: Daniel Almirall

Clinicians and researchers are increasingly interested in how best to individualize interventions. A dynamic treatment regime (DTR) is a sequence of pre-specified decision rules which guide the delivery of a course of treatments that is tailored to the changing needs of the individual. The sequential multiple-assignment randomized trial (SMART) is a research tool that can be used to inform the construction of effective DTRs. We introduce sample size formulae for SMARTs in which the primary aim is to compare two embedded DTRs using a continuous longitudinal outcome collected at three timepoints throughout the study. The method is based on a longitudinal analysis that accounts for unique features of a SMART, including modeling constraints and the over/under-representation of different sequences of treatment among participants. We also discuss extensions to a general number of timepoints. We illustrate the method using ENGAGE, a SMART aimed at developing a DTR for re-engaging patients with alcohol and/or cocaine use disorders who have dropped out of treatment.

EO270 Room MAL B20 STATISTICAL GENOMICS AND MACHINE LEARNING Chair: Wei Pan

E1699: **Deep learning in neuro-imaging genetics**
Presenter: **Wei Pan**, University of Minnesota, United States

Several convolutional neural networks/deep learning algorithms are first applied to brain MRI data from the ADNI to extract low-to-high level imaging features, which are then used in downstream analyses of a genome-wide association study (GWAS) to detect genetic variants associated with Alzheimers diseases. We discuss both some promising preliminary results and challenges in our application.

E1740: **Efficient algorithms for resampling-based hypothesis testing in genomic data analysis**
Presenter: **Hui Jiang**, University of Michigan, United States

Resampling-based hypothesis testing procedures such as bootstrapping and permutations tests are widely used in genomic data analysis when the distribution of the test statistic is analytically intractable. However, these test procedures are often computationally intensive, especially when the dataset is large, the desired significant level is very small or there are many tests to perform, all of which are commonly encountered scenarios in modern genomic studies. We will discuss several computational methods for accelerating such testing procedures while having theoretical justification and empirical evidence of achieving substantial speedup and high accuracy. The methods will be demonstrated with both simulated and real data experiments in genomics.

E1960: **Incorporating a large number of functional annotations in penalized linear regression**
Presenter: **Hu Yang**, Central University of Finance and Economics, China
Co-authors: Wei Pan

Many approaches based on penalized linear regression (such as lasso, adaptive lasso, elastic net, network penalty, etc.) have been successfully applied to analyze sparse and high-dimensional genetic data, including variable selection to facilitate interpretation of associations between predictors and a response. However, these methods are unable to incorporate a large number of function annotations, which can be collected from various omic studies, offering useful prior information. A new penalized linear regression method is proposed to incorporate such information so that we can better select predictors and predict gene expression.

EO701 Room MAL B35 ESTIMATION AND PREDICTION USING TIME-TO-EVENT DATA Chair: Feng-Chang Lin

E0337: **Causal inference for recurrent event data using pseudo-observations**
Presenter: **Chien-Lin Su**, McGill University, Canada
Co-authors: Robert Platt, Jean-Francois Plante

Recurrent event data are commonly encountered in observational studies where each subject may experience a particular event repeatedly over time. We aim to compare cumulative rate functions of two groups when treatment assignment may depend on the unbalanced distribution of confounders. Based on pseudo-observations, several estimators including inverse probability of treatment weighting estimator, regression model-based estimators and doubly robust estimators are proposed to adjust for the confounding effects. The proposed marginal regression estimator based on pseudo-observations is shown to be consistent and asymptotically normal. A bootstrap approach is proposed for the variance estimation of the proposed estimators. Model diagnostic plots of residuals are presented to assess the goodness-of-fit for the proposed regression models. A family of adjusted two-sample pseudo-score tests is proposed to compare group differences of cumulative rate functions. Simulation studies are conducted to assess finite sample performance of the proposed method. The proposed technique is demonstrated through an application to a hospital readmission data set.

E1774: **Analysis of cyclic recurrent event data with multiple event types**
Presenter: **Feng-Chang Lin**, University of North Carolina at Chapel Hill, United States
Co-authors: Chien-Lin Su

Recurrent event data frequently arise in practice, and in some cases, the event process has cyclic or periodic components. We propose a semiparametric rate model with multiple event types that have such features. Generalized estimating equations are used for the estimation of regression coefficients after profiling the baseline rate function with a fully nonparametric estimator. The proposed estimators are shown to be consistent and asymptotically Gaussian. Their finite-sample behavior is assessed through simulation experiments. The predictability of the model with and without the cyclic component is also compared. With the cyclic component, our model improves the predictability of a conventional model without the cyclic feature. Data on recurrent fire alarms in Blenheim, New Zealand, are used for illustration purposes.

E1990: **Transformation model estimation of survival under dependent truncation and independent censoring**
Presenter: **Sy Han Chiou**, University of Texas at Dallas, United States
Co-authors: Jing Qian, Rebecca Betensky, Matthew Austin
Truncation is a mechanism that permits observation of selected subjects from a source population; subjects are excluded if their event times are not contained within subject-specific intervals. Standard survival analysis methods for estimation of the distribution of the event time require quasi-independence of failure and truncation. When quasi-independence does not hold, alternative estimation procedures are required; currently, there is a copula model approach that makes strong modeling assumptions, and a transformation model approach that does not allow for right censoring. We extend the transformation model approach to accommodate right censoring. We propose a regression diagnostic for assessment of model fit. We evaluate the proposed transformation model in simulations and apply it to the National Alzheimers Coordinating Centers autopsy cohort study, and an AIDS incubation study. Our methods are publicly available in an R package, tranSurv.

EO602 Room MAL B36 OPTIMAL TRANSPORT AND STATISTICS Chair: Quentin Berthet

E1230: **Private learning and regularized optimal transport**
Presenter: **Vianney Perchet**, ENSAE & Criteo AI Lab, France
Private data are valuable either by remaining private (for instance if they are sensitive) or, on the other hand, by being used publicly to increase some utility. These two objectives are antagonistic and leaking data might be more rewarding than concealing them. Unlike classical concepts of privacy that focus on the first point, we consider instead agents that optimize a natural trade-off between both objectives. We formalize this as an optimization problem where the objective mapping is regularized by the amount of information leaked by the agent into the system (measured as a divergence between the prior and posterior on the private data). Quite surprisingly, when combined with the entropic regularization, the Sinkhorn divergence naturally emerges in the optimization objective, making it efficiently solvable. We apply these techniques to preserve some privacy in online repeated auctions.

E1343: **Maximum mean discrepancy gradient flow**
Presenter: **Anna Korba**, Gatsby Unit UCL, United Kingdom
A Wasserstein gradient flow of the maximum mean discrepancy (mmd) is constructed, and its convergence properties are studied. The MMD is an integral probability metric defined for a reproducing kernel Hilbert space (rkhs), and serves as a metric on probability measures for a sufficiently rich RKHS. We obtain conditions for convergence of the gradient flow towards a global optimum, that can be related to particle transport when optimizing neural networks. We also propose a way to regularize this MMD flow, based on an injection of noise in the gradient. This algorithmic fix comes with theoretical and empirical evidence. The practical implementation of the flow is straightforward, since both the MMD and its gradient have simple closed-form expressions, which can be easily estimated with samples.

E1389: **Sinkhorn divergences: Bridging the gap between optimal transport and MMD**
Presenter: **Aude Genevay**, MIT, United States
Sinkhorn Divergences, based on entropy-regularized OT, were first introduced as a solution to the computational burden of OT. However, this family of losses actually interpolates between OT (no regularization) and MMD (infinite regularization). This interpolation property is also true in terms of sample complexity, and thus regularizing OT breaks its curse of dimensionality. We will illustrate these theoretical claims on a set of learning problems like learning a distribution from samples.

EO466 Room Bloomsbury STATISTICAL MODELS FOR FINANCIAL DISTRESS Chair: Marialuisa Restaino

E0409: **Grabit: Gradient tree-boosted Tobit models for default prediction**
Presenter: **Fabio Sigrist**, Lucerne University of Applied Sciences, Switzerland
A frequent problem in binary classification, and in particular in default prediction, is class imbalance between a minority and a majority class such as defaults and non-defaults. We show how this issue can be alleviated by using a tree-boosted Tobit model in cases where there is auxiliary data for the non-default events that is related to the default mechanism. For instance, such auxiliary data can consist of number of days of delay by which loans were paid back, stock returns, rating changes, or distance to default measures. We apply our proposed model for predicting defaults on loans made to Swiss small and medium-sized enterprises and obtain a large improvement in predictive performance compared to other state-of-the-art approaches.

E0962: **Spatial dependence and localization in bankruptcy prediction: A comparative analysis on Italian manufacturing firms**
Presenter: **Maria Simona Andreano**, Universitas Mercatorum, Italy
Co-authors: Roberto Benedetti, Federica Piersimoni, Andrea Mazzitelli
The interest in prediction of firms bankruptcy is increased in recent years, when recession has sharply raised the number of distressed manufacturing firms. Although numerous studies focus on this topic and several attempts to provide a solution for the problem, predicting financial firms stress is not a trivial task. The most popular parametric models applied by bankruptcy researchers are the Logit and Probit, where failure is seen as a dichotomic event, whereas recently, failure prediction methods moved to more comprehensive machine learning techniques. The space and the location of firms have been considered a decisive factor in many fields of business-related research, however they are rarely applied in bankruptcy analysis. A spatial econometric methodology is applied to evaluate the effect of geographical location on the probability of business failure. Moreover, spatial dependence is included in machine learning techniques through an Iterated Conditional Modes (ICM) algorithm, originally introduced on image processing, based on non-degenerate Markov random fields. Non-spatial and spatial models are applied on about 12.000 manufacturing firms, located in Central Italy. Different evaluation metrics are selected to compare the performance of the various approaches. Our application shows that spatial contagion effects are an important issue when modelling bankruptcy probability and spatial models outperform classical ones.

E1497: **Nonlinear loss aversion and portfolio optimization**
Presenter: **Anna Maria Fiori**, University of Milano-Bicocca, Italy
Co-authors: Alessandro Avellone, Ilaria Foroni
The interest for loss averse preferences in portfolio optimization has been recently rekindled by increased financial market instability and by failures of various pension funds, also reported in the mainstream media. A number of papers have shown that loss averse portfolio theory can be made consistent with decision-theoretic models of choice under uncertainty and deliver superior performance relative to traditional Mean-Variance analysis. With a few notable exceptions, the majority of these papers have focused on linear or power loss aversion (LA), solving the optimal asset allocation problem either by linear programming or by Monte Carlo simulation for very small portfolios. The proposal puts forth a methodology that can be used to derive optimal asset allocation rules for general forms of nonlinear LA. The portfolio problem is solved by a new algorithm based on Particle Swarm Optimization (PSO), which leads to computationally efficient solutions even when a high number of assets is considered. Our variant of PSO permits the inclusion of real-world constraints (e.g., buy-in thresholds, cardinality) through a repair strategy that identifies unfeasible solutions and brings them back into the feasible domain of the investment problem. An empirical study is conducted to assess the impact of behavioral parameters and to evaluate the effective ability of LA portfolios to increase investor protection in adverse market conditions.

EO687 Room G11 INNOVATIVE STATISTICAL METHODS FOR META-ANALYSIS Chair: Lifeng Lin

E0471: **Hybrid test for publication bias in meta-analysis**
Presenter: **Lifeng Lin**, Florida State University, United States

Assessing publication bias is a critical procedure in meta-analyses for rating the synthesized overall evidence. Many statistical tests have been proposed to detect publication bias. However, they often make dramatically different assumptions about the cause of publication bias; therefore, they are usually powerful only in certain cases that support their particular assumptions, while their powers may be fairly low in many other cases. Although several simulation studies have been conducted to compare different tests' powers under various situations, it is infeasible to justify the exact mechanism of publication bias in a real-world meta-analysis and thus select the optimal publication bias test. We propose a hybrid test for publication bias by synthesizing various tests and incorporating their benefits, so that it maintains relatively high powers across various mechanisms of publication bias. The superior performance of the proposed hybrid test is illustrated using simulation studies and three real-world meta-analyses with different effect sizes. It is compared with many existing methods.

E0472: **Data-dynamic synthesis of historical information through network meta-analysis**
Presenter: **Jing Zhang**, University of Maryland, United States

Data-adaptive borrowing of historical information according to the consistency between the historical information and the new experimental data is gaining popularity in Bayesian clinical trial designs. It resolves the problems of reckless borrowing such as larger biases, higher type I error, and a lengthier and costlier trial, especially when prior-data conflict appears. We propose a novel network-meta-analytic-predictive prior (NMAPP) method by incorporating a network meta-analysis element in the synthesis of historical information. Unlike the existing method where only historical information of a single arm (usually the control group) is synthesized, the proposed method forms a prior using a network meta-analysis of multiple treatments from historical trials. Advantages of the proposed NMAPP method include that it (1) facilitates the design of multiple-arm trials; (2) avoids extracting single-arm information from randomized controlled trials; and (3) gains statistical efficiency thus further reduces sample size, cost, time and ethical hazard. Multi-component mixtures of conjugate priors are used as approximations to solve the problem of analytic unavailability. This mixture gains robustness and offers data-driven borrowing, and the conjugacy eases the posterior calculations. We illustrated the proposed methodology with two case studies. Simulation studies were conducted to evaluate the proposed method and to compare it to the existing method.

E0474: **A Bayesian hierarchical CACE model accounting for incomplete noncompliance data in meta-analysis**
Presenter: **Haitao Chu**, University of Minnesota School of Public Health, United States
Co-authors: Jincheng Zhou, James Hodges

Noncompliance to assigned treatments is a common challenge in the analysis and interpretation of a randomized clinical trial (RCT). One approach to handle noncompliance is to estimate the complier-average causal effect (CACE) using the principal stratification framework, where CACE measures the impact of an intervention in the subgroup of the population that complies with its assigned treatment. When non-compliance data are reported in each trial, intuitively one can implement a two-step approach (i.e., first, estimating CACE for each study, and then combining them using a fixed-effect or random effects model) to estimate the population-averaged CACE in a meta-analysis. However, it is common that some trials do not report noncompliance data. The two-step approach can be less efficient and potentially biased as trials with incomplete noncompliance data are excluded. We propose a flexible Bayesian hierarchical CACE framework to simultaneously account for heterogeneous and incomplete noncompliance data in a meta-analysis of RCTs. The performance of the proposed method is evaluated by extensive simulations, and an example of a meta-analysis estimating the CACE of epidural analgesia on cesarean section, in which only 10 out of 27 studies reported complete noncompliance data.

EO767 Room G21A BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS II Chair: Miguel Gonzalez Velasco

E1541: **Numerical schemes and algorithms for branching processes models in cancer**
Presenter: **Maroussia Slavtchova-Bojkova**, Sofia University, Bulgaria
Co-authors: Kaloyan Vitanov

A special class of reducible multi-type branching processes in continuous time is proposed as a powerful tool for studying the mutations in cancer cell populations. This model turns out to be useful for studying the dynamics of the number of different types of cells, which due to a small reproduction ratio are fated to become extinct. However, mutations occurring during the reproduction process, may lead to the appearance of a new type of cells that may escape extinction. We were deriving the limit distributions of the numbers of mutations of the escape type up to time t and in the whole process. A cell of the mutation type, which leads possibly to the beginning of a lineage, that will never become extinct is called successful mutant. These asymptotic results are used for developing numerical schemes and algorithms implemented in Python via the NumPy package for approximate calculation of the corresponding quantities. In conclusion, our conjecture is that this methodology can be advantageous in revealing the role of the lifespan distribution of the cancer cells in the context of cancer disease evolution and other complex cell population systems, in general.

E1951: **Likelihood-free simulation methodologies for controlled branching processes**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Carmen Minuesa Abril

Controlled branching processes (CBPs) are a family of discrete-time stochastic processes which are appropriate to describe population dynamics. This model generalizes the standard branching process - the so-called Galton-Watson process. As in this latter process, each individual reproduces independently of the others and following the same distribution, referred as the offspring law. The novelty of the CBP lies in the presence of a mechanism establishing the number of individuals with reproductive capacity (progenitors) in each generation. Thus, the evolution of populations suffering from the existence of predators, populations of invasive species or different migratory movements can be modelled by using this branching process. The behaviour of these processes are determined by the parameters of the model associated with the offspring and control laws and in real situations those values are unknown. The purpose is to examine likelihood-free simulation methodologies to obtain Bayesian inference for the main parameters of interest. These methodologies enable to approximate the posterior distribution of the parameters of interest satisfactorily without explicit likelihood calculations. In particular, we examine sequential Monte Carlo methods.

E1952: **Bayesian inference for a multitype two-sex branching process for X-linked recessive disorders**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Cristina Gutierrez Perez, Alicia Leon Naranjo, Rodrigo Martinez Quintana

Recently, a multitype two-sex branching process has been introduced for describing the evolution of the number of individuals carrying the alleles, R and r, of a gene linked to X chromosome. The R allele is considered dominant and the r allele is supposed to be recessive and defective, responsible of a disorder (hemophilia, red-green color blindness or the Duchenne and Becker's muscular dystrophies are examples of these diseases). For this model we investigate the estimation of its main parameters from a Bayesian standpoint. Concretely, we apply the Approximate Bayesian Computation (ABC) methodology to approximate its posterior distributions. The accuracy of the procedure is illustrated and discussed by way of simulated examples developed with R.

EO124 Room MAL G13 DOUBLY STOCHASTIC COUNTING PROCESSES Chair: Paula Bouzas

E0599: **Doubly stochastic Poisson processes in hydrological modelling**
Presenter: **Nadarajah Ramesh**, University of Greenwich, United Kingdom
The doubly stochastic Poisson process (DSPP) provides a rich class of clustered point process models that can be utilised in rainfall modelling. The purpose is to provide an overview of recent development on models constructed from this class of stochastic point processes and presents the results when they are used to model rainfall collected in different forms. When the rainfall is recorded in the form of rainfall bucket tip time series a class of DSPP models can be constructed whereby the arrival pattern of bucket tip times is viewed as a DSPP whose rate of occurrence varies according to a Markov process. As the likelihood function for this process can be calculated, the maximum likelihood methods can be employed to estimate the parameters. More physically appealing models from the DSPP can be developed, by attaching a pulse or a cluster of pulses to each rain cell, for rainfall collected in the form of accumulated rainfall in discrete intervals of fixed length. Different types of pulses can be employed to extend this model. These models are used to model hourly and sub-hourly rainfall data. The results of our analyses suggest that the proposed class of stochastic models provides useful tools in hydrological modelling.

E0608: **Multivariate Poisson processes with random effects to model spatial dependence**
Presenter: **Ana C Cebrian**, University of Zaragoza, Spain
Co-authors: Jesus Asin
Modeling the occurrence of events in many real problems related involves several Poisson processes. Frequently, these processes are dependent, and this feature should be considered when modeling. A common example of this situation is the spatial dependence appearing between the occurrence of events in different locations. Dependence between Poisson processes can be captured by allowing the intensities of the marginal models to be a function of common covariates. However, in many cases, adequate variables are not available, or the existing dependence is not totally captured by them. In those cases, we propose a multivariate vector of Poisson processes with random effects, where the intensities of the marginal processes are modeled as a function of covariates plus a common random effect. Under quite mild conditions, an integrated nested Laplace approximation can be used to estimate this model. This approach is used to model the dependence between the occurrence of extreme heat events (EHEs) in several locations in the North-East of Spain (Aragon). The pairwise dependence is analyzed and a multivariate Poisson process with random effects is used to obtain a joint model for the occurrence of EHEs.

E1214: **Doubly stochastic point processes in time and space as a model for photoactivated localization microscopy data**
Presenter: **Louis Gammelgaard Jensen**, Aarhus University, Denmark
Co-authors: Ute Hahn
Photoactivated localization microscopy (PALM) is a revolutionary technique that promises to localize single molecules, and increases resolution from 300nm to 20nm. The method uses photoactivatable fluorophores attached to the molecules of interest. An image is obtained from fluorescence photons that are registered in a video. Only few molecules are activated by a laser beam at a time, thus only few photons are registered in each video frame. This makes it possible to distinguish between photons coming from different molecules. Reconstruction of single molecule positions in PALM is crucially based on the assumption that the fluorophores are activated independently of each other. It is known though, that several photons can be emitted from the same molecule over time, an effect called photoblinking. This may lead to artifacts in the interpretation of the raw data. We set up a doubly stochastic model for the data, with a latent spatial point process representing the molecules. This process serves as parent process for a space-time cluster process, reflecting photoblinking. We present statistical methods for estimating the model parameters, allowing us to separate contributions from molecules and blinking artifacts.

EO068 Room MAL G14 PROJECTION PURSUIT: THEORY Chair: Nicola Loperfido

E0422: **Multivariate kurtosis, projection pursuit and tensor eigenvectors: A triangulation**
Presenter: **Nicola Loperfido**, University of Urbino, Italy
Kurtosis-based projection pursuit looks for interesting data structures by means of projections with either maximal or minimal kurtosis. The projecting directions and the corresponding kurtoses coincide with the tensor eigenvectors and tensor eigenvalues of the fourth standardized moment, regarded as a fourth-order, real and symmetric tensor. Their properties are investigated both in the general case and for some statistical models, as for examples finite mixtures and hidden truncation models. Kurtosis-based projection pursuit is closely related to outlier detection, cluster analysis, independent component analysis, normality testing and portfolio selection. Its practical relevance is illustrated with well-known datasets: the Iris dataset, the Crab dataset and the Australian Athletes dataset.

E0870: **Affine vs orthogonal equivariance in multivariate analysis**
Presenter: **John Kent**, University of Leeds, United Kingdom
Many of the classic methods in multivariate analysis (e.g. Hotelling's T^2, Fisher's linear discriminant analysis, MANOVA, canonical correlation analysis, independent component analysis) are affine equivariant. That is, they give essentially the same answer if the data undergo an affine transformation. However, other methods, especially those that can be used for high-dimensional data, are only orthogonally equivariant. Examples include PCA, ridge regression, PLS, k-means clustering, support vector machines, and projection pursuit. The restriction to orthogonal equivariance is an example of regularization. Some form of regularization is essential if a statistical method is to be applicable in high-dimensional problems. The focus will be on what is gained and what is lost by this restriction. In addition, comparisons will be made to variable-based methods of regularization, such as decision trees and LASSO-type penalties.

E1109: **Limitations of projection pursuit methods for high dimensional data**
Presenter: **Joao Antonio Branco**, CEMAT Instituto Superior Tecnico Universidade de Lisboa, Portugal
Co-authors: Ana Maria Pires
Any multivariate statistical method which can be defined as "find a low-dimensional orthogonal projection of the data cloud such that a given statistic of the projected data is optimal", is a projection pursuit method. Two famous examples are principal components and linear discriminant analysis. The advantage of the formulation as a projection pursuit method is that it offers an easy way to generalize the original method and to lift properties of univariate methods (assuming low = 1) to multivariate methods. For example, using appropriate robust univariate estimators one can easily produce robust principal components or robust linear discriminant functions. For some time it was thought that projection pursuit could bypass the curse of dimensionality. Unfortunately that is not the case. Using some recent results from high-dimensional geometry we will uncover some of the limitations of projection pursuit methods for high dimensional data and discuss possible alleviating solutions.

EO326 Room MAL G15 BAYESIAN SPATIAL MODELLING Chair: Silvia Liverani

E0820: **Detecting life expectancy anomalies in England using a Bayesian hierarchical model**
Presenter: **Areti Boulieri**, Imperial College London, United Kingdom
Co-authors: Marta Blangiardo
In England, life expectancy has shown a steady increase over many years, however these improvements have recently started to slow down considerably. The aim is to investigate the changes in life expectancy in England over time and across its local authorities, and to identify local authorities with unusual time trends that might help with hypothesis generation and point to emerging risk factors. We analyse mortality count data

in England for females at the local authority level (324 areas), from 2001 to 2016 (17 years), and by age group, assuming 19 age groups of 5 year bands. We develop a statistical model within the Bayesian hierarchical framework that accounts for spatial, temporal, and age effects, as well as for pairwise interactions. The space-time interaction parameter is used to detect areas whose time trends deviate from the national one. The detection rule that we specify focuses on areas that are detected as unusual over the last 5 years of the time period 2013-2017. The model is implemented in Integrated Nested Laplace Approximations (INLA). We found roughly 40 areas to be highlighted as unusual under the model, following a different time trend in the mortality rates compared to the national trend.

E1095: **Bayesian statistical machine learning methods for mapping health and development metrics using big data**
Presenter: **Chigozie Utazi**, University of Southampton, United Kingdom

Recent technological developments such as the use of Global Positioning Systems (GPS) in surveys and data capture have resulted in unprecedented scales and frequencies at which data are collected. In the global health and development (H&D) arena, this development has led to a rapid increase in the availability of and access to geo-referenced information for monitoring H&D metrics. Many nationally representative household surveys are now geo-coded and increasing numbers of survey clusters typically geo-coded enumeration areas - are being included in more recent surveys that can enhance the estimation of H&D indicators at finer scales than administrative-level one areas. Mapping spatial big data using Bayesian model-based geostatistical approaches poses computational challenges. We develop statistical machine learning methods for high resolution mapping of H&D indicators using spatial big data. The methodology is an amalgamation of subsampling and ensemble approaches for fitting spatial generalized linear models. We consider random and stratified subsampling and model ensembles analogous to Bayesian model averaging. The Bayesian method is implemented using the INLA-SPDE approach and applied to mapping vaccination coverage using Demographic and Health Survey data. The output maps highlight significant heterogeneities in coverage levels which are indispensable for program planning and implementation.

E1171: **Sparsity priors in spatial models**
Presenter: **Peter Congdon**, QMUL, United Kingdom

Sparsity inducing priors are widely used in regression analysis, and seek dimensionality reduction to distinguish significant predictors and avoid unnecessarily complex models. An alternative to sparsity induction are discrete mixtures, such as spike and slab priors. These ideas extend to selection of random effects, either i.i.d or structured (e.g. spatially structured) with the goal of distinguishing significant random effects. We consider the use of sparsity inducing priors, and of global-local shrinkage, to distinguish areas with distinct spatial effects using sparsity priors applied to a spatial intercept. Sparsity priors are also considered to distinguish areas with average predictor effects from areas with amplified or diminished predictor effect because the response-predictor pattern is distinct from that of most areas. The operation and utility of this approach is demonstrated using simulated data, and a real application to diabetes related deaths in New York counties.

EO362 Room MAL G16 ADVANCES IN BAYESIAN MODELLING Chair: Maria De Iorio

E1914: **A Bayesian approach to the joint analysis of multi-type image-based and coordinate-based neuroimaging meta-analysis data**
Presenter: **Silvia Montagna**, University of Turin, Italy
Co-authors: Thomas Nichols, Timothy Johnson

As the popularity of functional MRI (fMRI) has grown exponentially over the years, so does the need to aggregate and summarise different fMRI studies via meta-analysis. Neuroimaging meta-analysis is used to 1) identify areas of consistent activation; and 2) build a predictive model of task type or cognitive process for new studies (reverse inference). Currently, two types of meta-analyses are possible. Namely, coordinate-based meta-analyses (CBMA); when data from different studies are available only as peak activation coordinates (foci) in a three dimensional coordinate system. And, image-based meta-analyses (IBMA); when the statistical parametric maps resulting from a group-level analysis (voxel-level data) are shared. We propose a Bayesian hierarchical model for neuroimaging meta-analysis which allows for the joint modelling of CBMA and IBMA data, whilst simultaneously addressing the two aims above. Specifically, we build a spatial process for voxel-level (IBMA) data and a spatial point process model for point-pattern foci-based (CBMA) data, then combine the two component models via a latent factor framework that allows for the borrowing of information across the different studies. We apply our methodology to a neuroimaging meta-analysis dataset of pain and emotions studies.

E2007: **A Bayesian decision-theoretic design for a treatment-selection biomarker**
Presenter: **Gary Rosner**, Johns Hopkins University, United States
Co-authors: Zheyu Wang, Chenguang Wang

Drug development, particularly in oncology, often focuses on developing therapies that target molecular pathways in an attempt to disrupt disease processes or alleviate symptoms. Successful drug development often relies on the ability to select appropriate patient subpopulations that are more likely to respond to the treatment. As a result, clinical studies of these targeted agents often include biomarker assessment, particularly early studies of the treatment's safety and activity. We propose a two-stage design based on a Bayesian decision-theoretic approach to achieve the dual aim of biomarker subgroup selection and efficacy demonstration. Stage 1 enrolls patients regardless of their biomarker values. An analysis at the end of stage 1 identifies a biomarker threshold based on Stage 1 data and any external information that may be available. The second stage enrolls either all patients or a biomarker-defined subset of patients, depending on the interim analysis results. We will discuss the design and its characteristics in light of the particular challenges and opportunities for clinical trial design of targeted therapies.

E1851: **Modelling ethnic differences in metabolic associations via Bayesian nonparametric processes**
Presenter: **Marco Molinari**, University College London, United Kingdom

A novel approach is proposed for the estimation of multiple Gaussian Graphical Models (GGMs) to analyse patterns of association among a set of metabolites, under different conditions. Our motivating application is the Southall And Brent REvisited (SABRE) study, a tri-ethnic cohort study conducted in the UK. We are interested in identifying potential ethnic differences in metabolite levels and associations, with the aim of gaining a better understanding of different risk of cardio-metabolic disorders across ethnicities. We model the relationship between a set of metabolites and a set of covariates through a Sparse Seemingly Unrelated Regressions model and we use GGMs to represent the conditional dependence structure among metabolites. We specify a Dependent Generalised Dirichlet Process prior on the edge inclusion probabilities to borrow strength across groups and we adopt the Horseshoe prior to identify important biomarkers. Inference is performed via Markov Chain Monte Carlo (MCMC).

EO831 Room CLO 101 RECENT DEVELOPMENTS ON DATA DEPTH AND ITS APPLICATIONS Chair: Pavlo Mozharovskyi

E0619: **The halfspace depth characterization problem**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic

The halfspace depth is an inferential tool that aims to generalize quantiles to multivariate datasets. It has been long conjectured that, just as for the usual quantiles, there is a one-to-one relation between all Borel probability measures, and all possible depth surfaces. We answer this conjecture in the negative. That suggests an interesting open problem of characterizing those probability measures that possess a unique depth. A complete solution to this problem would have far-reaching implications, not only in the theory of multivariate statistics.

E0709: **The area of the convex hull of sampled curves: A robust functional statistical depth measure**
Presenter: **Guillaume Staerman**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi, Stephan Clemencon

With the increasing industrial digitalization contemporary data are often present in the form of temporal series or functions. Out of existing statistical tools for functional data analysis, statistical data depth distinguishes by its non-parametric nature and robustness. Having undergone theoretical and computational developments in the recent decades, it has proven to be of particular use in functional spaces. Nevertheless, most of the existing functional depths share a common feature of treating evaluations for different arguments independently of each other, and by that may possess certain insensitivity to the shape changes. We propose a notion of functional depth based on the area of the convex hull of the functions' graphs. This approach allows for capturing gradual departures from centrality, even beyond the envelope of the data, but additionally provides a possibility to simultaneously consider functional evaluations for multiple arguments. We discuss the practical relevance of commonly imposed axioms on functional depths and their satisfaction by the proposed notion, and construct an efficient estimation algorithm. An extension to generic geometric transformations is also suggested, as well as a generalization to multivariate functional data. Simulation and real data studies demonstrate exploratory properties of the developed depth function. In particular, its application for functional anomaly detection is advantageous.

E0815: **Imputation of missing values by pseudo-marginal simulated annealing algorithm using depth statistics**
Presenter: **Kimsy Tor**, Telecom ParisTech, France
Co-authors: Randal Douc, Pavlo Mozharovskyi, Francois Roueff

The problem of missing data is peculiar to many applications and often impedes statistical analysis. One of the universal ways to deal with missing values is their imputation prior to employment of the statistical method in question. Since model-based imputation approaches are limited to particular data-generating processes, non-parametric imputation methods have been proposed in the literature. These however induce high computational cost due to parameter tuning and suffer from the curse of dimensionality. We propose a new imputation method based on the non-parametric robust measure of centrality called data depth. The depth statistics is used to generate a family of models which are then fitted simultaneously using the pseudo-marginal simulated annealing algorithm. The created models are chosen to be simple enough to be easily fitted and flexible enough to suit many applications. Multiple imputed data sets constitute direct output of the procedure while single imputation can be obtained by their averaging, which makes the method suitable for both estimation and inference. Simulation and real-data studies illustrate competitive performance of the proposed approach.

EO490 Room MAL 152 NEW ADVANCES IN NONPARAMETRIC BAYESIAN METHODS Chair: Guanyu Hu

E0398: **A Bayesian semiparametric approach for spatial nonhomogeneous Poisson process with applications**
Presenter: **Jieying Jiao**, University of Connecticut, United States
Co-authors: Guanyu Hu, Jun Yan

Spatial point pattern data are routinely encountered in various fields such as seismology, ecology, environmental science, and epidemiology. Building a flexible regression model for spatial point process is an important task in order to reveal data's spatial pattern and relationships with various factors. We propose a Bayesian semiparametric regression model for spatial Poisson point process data based on powered Chinese resturant process. Further, we allow variable selection through the spike-slab prior. An efficient Markov chain Monte Carlo (MCMC) algorithm is developed for the proposed methods, followed with an extensive simulation studies to evaluate the empirical performance. The proposed methods are further applied to the analysis of the Forest of Barro Colorado Island (BCI) data.

E0324: **Nonparametric Bayesian functional clustering for breast cancer disparities**
Presenter: **Wenyu Gao**, Virginia Tech, United States
Co-authors: Wonil Nam, Inyoung Kim, Wei Zhou

It has been found that different incidence and mortality rates for breast cancer exist among various racial populations. For instance, Caucasian women are more likely to develop breast cancer than African American women. To study these disparities, surface-enhanced Raman spectroscopy (SERS) has been conducted to provide biomolecular fingerprint information. Extracellular SERS signals from each cell type were measured by a practical high-performance SERS device. However, large intraclass variations exist due to cellular and additional cancerous heterogeneity. To study the differences between two types of triple negative breast cancer cell lines at the molecular level, we performed clustering analyses on the massive nonlinear curves of signals versus Raman shifts. We propose a nonparametric Bayesian functional clustering method via Weighted Dirichlet Process Mixture (WDPM) modeling, which clusters automatically and determines correct number of clusters. Based on our analyses, we identify that the clustering behaviors vary across different racial groups. This clustering information will be useful to further investigate health disparities among groups.

E0456: **Bayesian spatial homogeneity pursuit for survival data**
Presenter: **Lijiang Geng**, University of Connecticut, United States
Co-authors: Guanyu Hu

A new Bayesian spatial homogeneity pursuit method is proposed for survival data under the Cox proportional hazards model, to detect spatially clustered patterns in the associations between hazard rates and covariates. Specially, regression coefficients and baseline hazard are assumed to have spatial homogeneity over space. To capture the homogeneity in regression coefficients and baseline hazards, we develop a geographically weighted Chinese restaurant process prior for them. An efficient Markov chain Monte Carlo (MCMC) algorithm is designed to estimate the clustered coefficients and baseline hazards and their uncertainty measures simultaneously. Extensive simulations are conducted to evaluate the empirical performance of the proposed models. Finally, we illustrate the performance of the model with a real data analysis of respiration cancer in Louisiana.

EO154 Room MAL 252 MULTIVARIATE HIGH-DIMENSIONAL STATISTICAL LEARNING Chair: Teng Zhang

E0535: **Conformal prediction for exponential families and generalized linear models**
Presenter: **Daniel Eck**, University of Illinois, United States

Conformal prediction methods construct prediction regions for iid data that are valid in finite samples. Distribution-free conformal prediction methods have been proposed for regression. Generalized linear models (GLMs) are a widely used class of regression models, and researchers often seek predictions from fitted GLMs. We provide a parametric conformal prediction region for GLMs that possesses finite sample validity and is asymptotically of minimal length when the model is correctly specified. This parametric conformal prediction region is asymptotically minimal at the $\sqrt{\log(n)/n}$ rate when the dimension d of the predictor is one or two, and converges at the $O\{(\log(n)/n)^{1/d}\}$ rate when $d > 2$. We develop a novel concentration inequality for maximum likelihood estimation in exponential families that induces these convergence rates. We analyze prediction region coverage properties, large-sample efficiency, and robustness properties of four methods for constructing conformal prediction intervals for GLMs: fully nonparametric kernel-based conformal, residual based conformal, normalized residual based conformal, and parametric conformal which uses the assumed GLM density as a conformity measure.

E0668: **Insights and algorithms for the multivariate square-root lasso**
Presenter: **Aaron Molstad**, University of Florida, United States

The multivariate square-root lasso is studied, which is a method for fitting the multivariate response linear regression model with dependent errors. This estimator minimizes the nuclear norm of the residual matrix plus a convex penalty. Unlike some existing methods for multivariate response linear regression, which require explicit estimates of the error covariance matrix or its inverse, the multivariate square-root lasso criterion implicitly

adapts to dependent errors and is convex. To justify the use of this estimator, we establish an error bound which illustrates that like the univariate square-root lasso, the multivariate square-root lasso is pivotal with respect to the unknown error covariance matrix. Based on our theory, we propose a simple tuning approach which requires fitting the model for only a single value of the tuning parameter, i.e., does not require cross-validation. We propose two algorithms to compute the estimator: a prox-linear alternating direction method of multipliers algorithm, and a fast first order algorithm which can be applied in special cases. In both simulation studies and a real data application, we show that the multivariate square-root lasso can outperform more computationally intensive methods which estimate both the regression coefficient matrix and error precision matrix.

E1687: **Spatio-temporal models for big multinomial data using the conditional multivariate logit-beta distribution**
Presenter: **Jonathan Bradley**, Florida State University, United States

A Bayesian approach is introduced for analyzing high-dimensional multinomial data that are recorded over space and time. In particular, the proportions associated with multinomial data are assumed to have a logit link to a latent spatio-temporal mixed effects model. This strategy allows for nonstationarity covariances in both space and time, asymmetry covariances, and dimension reduction. We also use the conditional multivariate logit-beta distribution, which leads to conjugate full-conditional distributions for use in a collapsed Gibbs sampler. Additionally, we provide methodological developments including (but not limited to): the derivation of the associated full-conditional distributions, a relationship with a latent Gaussian process model, and the stability of the non-stationary vector autoregressive model. We illustrate our model through simulations and through a demonstration with public-use quarterly workforce indicators data from the longitudinal employer household dynamics program of the US Census Bureau.

EO518 Room MAL 253 DATA CONFIDENTIALITY FOR FREQUENCY TABLES Chair: Stefano Favaro

E0236: **Locally private Bayesian inference for Poisson factorization models**
Presenter: **Aaron Schein**, Columbia University, United States
Co-authors: Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, Hanna Wallach

A general method is presented for privacy-preserving Bayesian inference in Poisson factorization, a broad class of models that includes some of the most widely used models in the social sciences. Our method satisfies limited precision local privacy, a generalization of local differential privacy, which we introduce to formulate privacy guarantees appropriate for sparse count data. We develop an MCMC algorithm that approximates the locally private posterior over model parameters given data that has been locally privatized by the geometric mechanism. Our solution is based on two insights: 1) a novel reinterpretation of the geometric mechanism in terms of the Skellam distribution and 2) a general theorem that relates the Skellam to the Bessel distribution. We demonstrate our method in two case studies on real-world email data in which we show that our method consistently outperforms the commonly-used naive approach, obtaining higher quality topics in text and more accurate link prediction in networks. On some tasks, our privacy-preserving method even outperforms non-private inference which conditions on the true data.

E0352: **Privacy in data dissemination, differential privacy, and analysis of perturbed data**
Presenter: **Yosef Rinott**, The Hebrew University, Israel

Privacy issues that arise when an agency disseminates data will be briefly reviewed, along with some of the methods used by statisticians to assess the disclosure risk, and to decrease it. In general, such methods depend on scenarios regarding prior knowledge of potential intruders and the nature of the disseminated data. Differential Privacy is an approach that avoids much of the need to consider such scenarios, and guarantees a well-defined notion of privacy by adding noise with a known distribution to all released data. Some basic results on differential privacy and applications to the release of contingency tables will be discussed. In many cases, the released data appears like real data, and many researchers tend to analyze it without taking the noise distribution into account. Ongoing work on data analysis that takes the added noise into account, and the loss incurred by ignoring it will be discussed.

E1085: **Optimal nonparametric disclosure risk assessment**
Presenter: **Francesca Panero**, University of Oxford, United Kingdom
Co-authors: Stefano Favaro, Federico Camerlenghi, Zacharie Naulet

An original nonparametric estimator is presented for the number of unique individuals in a sample that are also unique in the population, a classical measure of disclosure risk for microdata files. This estimator is easy to derive, scalable to massive datasets and can be interpreted as empirical Bayesian. We prove that it is nearly optimal by showing that the limit of predictability of it, in terms of vanishing normalized mean squared error, matches asymptotically with the maximum possible value that can be achieved by any nonparametric estimator. In particular, for a sample of size n and a population of size $n+\lambda n, \lambda > 0$, we show that our estimator is optimal for λ growing not faster than the logarithm of n. This result answers a long standing question about the feasibility of nonparametric estimation of this problem under the only assumption of the Poisson abundance model.

EO116 Room SH349 STATISTICS FOR COMPLEX INFERENCE PROBLEMS IN DATA SCIENCE Chair: Shevaun Neupert

E0384: **On the calibration of Bayes factors**
Presenter: **Jan Hannig**, University of North Carolina at Chapel Hill, United States
Co-authors: Hari Iyer

Many computer programs and software systems used in the interpretation of forensic evidence have as their output Bayes factors also commonly referred to as likelihood ratios. For example, it is not unusual to see it reported that the DNA recovered at the crime scene is a million times more likely under the assumption that the defendant is a contributor to the crime stain than under the assumption that the defendant is not a contributor. We summarize existing approaches for examining the validity of likelihood ratio systems and discuss a new statistical methodology, based on generalized fiducial inference, for empirically examining the validity of such likelihood ratio assessments. Using data from ta number of sources, such as glass, paint and DNA evidence, we illustrate our approach by examining LR values calculated using standard approaches in forensic literature.

E1090: **An email-experiment to identify the effect of racial discrimination on legal assistance: A statistical approach**
Presenter: **Tirthankar Dasgupta**, Rutgers University, United States
Co-authors: Brian Libgober

The problem of conducting an email experiment is considered to study the probable effect of racial bias of clients on access to lawyers. The problem of discriminating between potential linear and non-linear effects of racial signal is formulated as a statistical inference problem with the objective of inferring about a parameter that determines the shape of a specific function. Various complexities associated with the design and analysis of this experiment are handled by applying a novel combination of rigorous, semi-rigorous and rudimentary statistical techniques. The actual experiment is performed with a population of lawyers in Florida.

E1811: **Recent developments in Bayesian multivariate one-way ANOVA models**
Presenter: **Dongchu Sun**, University of Nebraska-Lincoln, United States

The multivariate one-way ANOVA model is important in contemporary statistical theory and application. The model has an unknown overall mean and two unknown covariance matrices, the error covariance matrix and the random effects covariance matrix. We study this problem from the Bayesian perspective. Typically, independent prior distributions are assumed for the mean and each covariance matrix; that case is considered

herein, with the primary focus being the determination of when common objective priors yield proper posteriors. We study a new class of dependent priors called "commutative priors", motivated from three directions. First, there are problems where it is most natural to utilize a prior on the "signal to noise ratio" (here a matrix); second, one often tries for dimension reduction, and the commutative priors substantially reduce the dimension of the unknowns; third, the commutative priors have excellent computational features. Interestingly, the commutative prior is also a conjugate prior. Propriety and moment existence are derived for both the priors and their posteriors. Moreover, a new and computationally effective MCMC algorithm is developed for the proposed commutative priors. Simulation and real data analysis show the potential advantages of the commutative priors.

EG732 Room MAL 153 CONTRIBUTIONS IN MIXED MODELS Chair: Christopher McMahan

E1548: **Simultaneous inference for empirical the best predictor under generalized linear mixed models**
Presenter: **Katarzyna Reluga**, University of Geneva, Switzerland
Co-authors: Maria Jose Lombardia, Stefan Sperlich

Simultaneous inference is considered for the empirical best predictor under generalized linear mixed models. In particular, we propose a method to construct simultaneous prediction intervals (SPIs). To the best of our knowledge, SPIs have not been developed under this modelling framework. SPIs allow researchers and practitioners to carry out statistically valid multiple comparisons of all or several parameters of interest. Aforementioned analysis can be desirable within certain domains such as small area estimation, which is often applied in, among others, studies measuring poverty, policy-making or ecological and demographic projects. Moreover, we develop a multiple testing procedure employing a max-type statistic. We focus on the maximum likelihood based estimation. We provide some details regarding the area-level Poisson model. A proof of the asymptotic coverage probability of simultaneous bands is provided. The theoretical results are accompanied by an extensive simulation experiment and a data example. The latter reveals an advantage of SPIs in the simultaneous study of the estimators. On the other hand, in this situation, the cluster-wise confidence intervals do not account for the variability arising from the joint statements and may lead to completely erroneous conclusions.

E1653: **qape: R package to estimate prediction accuracy in mixed models based on bootstrap methods**
Presenter: **Alicja Wolny-Dominiak**, University of Economics in Katowice, Poland
Co-authors: Tomasz Zadlo

In order to compare relevance of two models, the coefficient of determination, residual variance or AIC are typically calculated. One of the most popular prediction accuracy measures is the mean squared error of prediction (MSEP) and its different modifications. The alternative is the quantile absolute prediction error (QAPE) reflects the relation between the magnitude of the error and the probability of its realization. The R package qape to estimate such errors is presented. The parametric and residual bootstrap procedures, as well as the double bootstrap, are implemented. The LMM, GLMM and GLM models are applied.

E1922: **Accurate likelihood inference on boundaries**
Presenter: **Soumaya Elkantassi**, Ecole Polytechnique Federale de Lausanne, Switzerland
Co-authors: Anthony Davison

Statistics used to test hypotheses concerning parameters on the boundary of their domain often have non-standard limiting distributions, and these may be poor approximations to finite-sample distributions of the test statistics even when the sample size is very large. A canonical example of such a situation is testing for a zero variance component, which is equivalent to testing whether a spline expansion is needed in semi-parametric regression. An approach is described to small-sample approximation in such settings, based on higher-order approximations and in particular the tangent exponential model. Numerical results show that the approach can give much improved approximations, even in small samples.

EG856 Room MAL 251 CONTRIBUTIONS IN COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Cristian Gatu

E1646: **The Almon M-estimator for the distributed lag models in the presence of outliers**
Presenter: **Aslam Muhammad**, Bahauddin Zakariya University, Pakistan
Co-authors: Abdul Majid

The Almon technique is widely used for estimation of the distributed lag model (DLM) to encounter the problems associated with the application of ordinary least squares (OLS) to this model. The Almon estimator (AE) may be sensitive to outliers in y-direction. The aim is to propose a robust estimator for parameters of the DLM when the data set contains outliers. The performance of the proposed estimator is evaluated through Monte Carlo simulations. The simulation results reveal an attractive performance of the proposed estimator in presence of outliers.

E2020: **Testing and modelling time series with time varying tails**
Presenter: **Dario Palumbo**, University of Cambridge, United Kingdom

The occurrence of extreme observations in a time series depends on the heaviness of the tails of its distribution. A dynamic conditional score model (DCS) is proposed for modelling dynamic shape parameters that govern the tail index. The model is based on the Generalised t family of conditional distributions, allowing for the presence of asymmetric tails and therefore the possibility of specifying different dynamics for the left and right tail indices. Both the convergence properties of the model and the implications of the used link functions are examined by simulations. In addition, the size and power properties of a new Lagrange multiplier test to detect the presence of dynamics in the tail index parameter are introduced and studied. The model is fitted to Equity Indices and Credit Default Swaps returns. It is found that the tail index for equities has dynamics driven mainly by the lower tail, whereas for Credit Default Swap the test identifies very persistent dynamics for both the tails. Finally the implications of dynamic tail indices for the estimated conditional distribution are assessed in terms of time-varying quantiles.

E2027: **Entropy-based criteria for multivariate association and omics network models**
Presenter: **Takoua Jendoubi**, Imperial College London, United Kingdom
Co-authors: Korbinian Strimmer

In the last twenty years, the parallel acquisition of high-throughput omics datasets has seen a tremendous boost pushing forward deeper understanding of biological functions and molecular mechanisms. This is commonly achieved by investigating the degree of co-expression between omics variables, which is often estimated using measures of association such as correlation coefficients. Despite the widespread use of these measures, they do not convey information about the entire multivariate system. On the other hand, multivariate measures of association generalizing correlations coefficients to two random vectors, such as the RV coefficient or the distance covariance (dCov) coefficient, do not explore pairwise contributions to total association between variables which may be crucial to uncover inter-omics interactions. To address these drawbacks, we propose to use vector correlation coefficient, which seems to be largely ignored, as an alternative to the RV and dCov coefficients. By entropy derivation we show that this approach is natural in the setting of latent-variable multivariate regression and probabilistic canonical correlation analysis. In addition, we show that this measure offers a decomposition property allowing to dissect the total association in order to construct a network with pairwise contributions. We illustrate our approach by analyzing both synthetic and publicly available omics data.

E1779: **Comparison of classes of generalized Hill estimators**
Presenter: **Frederico Caeiro**, NOVA.ID.FCT - Universidade Nova de Lisboa, Portugal
Co-authors: Ivette Gomes, Ivanilda Cabral

The focus is on the estimation of the extreme value index, the primary parameter of extreme events. For heavy tails, classical extreme value index

estimators, such as the Hill estimator, have usually a strong bias. Consequently those estimators are quite sensitive to the number of top order statistics used in the estimation. To improve the mean squared error of the aforementioned estimators, many alternative estimators have appeared in the literature. We analyse several generalizations of the Hill estimator. The aim is to study their non degenerate asymptotic behaviour and to compare them altogether.

CO861 Room G3 PREDICTIVE ACCURACY METHODS Chair: Emanuela Raffinetti

C0432: **Simple ways to interpret effects in modeling binary and ordinal data**
Presenter: **Alan Agresti**, University of Florida, United States

Probability-based effect measures for models for binary and ordinal response variables can be simpler to interpret than logistic and probit regression model parameters and their corresponding effect measures, such as odds ratios. For describing the effect of an explanatory variable while adjusting for others in modeling a binary response, it is sometimes possible to employ the identity and log link functions to generate simple effect measures. When such link functions are inappropriate, one can still construct analogous effect measures. For comparing groups that are levels of categorical explanatory variables or relevant values for quantitative explanatory variables, such measures can be based on average differences or ratios of the probability modeled. For quantitative explanatory variables, they can also be based on average instantaneous rates of change for the probability. Analogous measures are proposed for interpreting effects in models for ordinal responses and with nonlinear predictors, such as generalized additive models, and are illustrated with an example implemented with R software.

C0497: **A rank graduation measure to assess predictive accuracy**
Presenter: **Emanuela Raffinetti**, University of Milan, Italy
Co-authors: Paolo Giudici

A very key point in the application of statistical and machine learning methods in Artificial Intelligence (AI) is the evaluation of their predictive accuracy. This is because the "automatic" choice of an action crucially depends on the predictive scenario under which that action will be implemented. Machine learning and statistics have provided, over the years, a number of summary measures aimed at measuring predictive accuracy, such as the root mean squared error, and the area under the ROC curve. Note that most of them are response-specific, and none of them can be applied to all types of response. This can be a problem in a complex situation, with different types of responses and, more generally, for an Artificial Intelligence system whose evaluation criteria should be determined exogenously and not endogenously. The aim is to present a more general measure which can improve predictive accuracy assessment in highly complex situations. More precisely, the proposed measure, called Rank Graduation index, is based on the comparison between the observed and the predicted response variable ranks, as in ordinal response models, but using, rather than the ranks themselves, the actual values of the response variable corresponding to both ranks, as in continuous or 0/1 response models. In order to appreciate the RG features, an application to credit scoring is also considered.

C0517: **Evaluation of a hydro-economic forecasting system as a support tool for energy trading**
Presenter: **Nicola Di Marco**, Free University of Bozen/Bolzano, Italy
Co-authors: Francesco Ravazzolo, Maurizio Righetti

Real-time balancing of the electricity grid is increasingly challenging due to intermittent supply of energy produced by non-programmable energy sources, such as wind farms, solar, photovoltaic and run-of-the river power plants. Operators on electricity markets have therefore to be careful in scheduling the amount of energy and the bidding price in order to optimize the energy trading as well as to minimize electricity grid unbalancing. The aim is to help trading operators involved in hydro-power energy generation combining forecasts of Northern Italian hourly electricity prices and water inflow to a run-of-the-river power plant, installed in the South-Tyrol region, East-Northern Italy. Prices are predicted with an ARMA model augmented with demand and production information such as renewable energy resources and then combined with hourly water inflow predictions (a proxy for the energy production), which reliability is mostly affected by weather forecast uncertainties. The accuracy of the hydro-economic forecasts are evaluated through a back-test carried out over the period May-October 2019, where hydro-power production data are available. Results show promising performance of the proposed system, even though the accuracy changes over the year and depending on the statistical metric employed.

CO384 Room G4 PRICE DISCOVERY AND LIQUIDITY IN MODERN FINANCIAL MARKETS Chair: Mohammad Jahan-Parvar

C1331: **What makes HFTs tick?**
Presenter: **Alain Chaboud**, Federal Reserve Board, United States
Co-authors: Clara Vega, Avery Dao

The purpose is to study the impact that two trading rule changes in the interdealer spot foreign exchange market, a reduction in the tick size and a subsequent increase, had on the trading behavior of various types of market participants. We find that the most notable impact of the tick size reduction was a substantial increase in the liquidity demand of high-frequency traders (HFTs), not the decrease in their liquidity provision predicted by recent literature. We show that this change in behavior was linked to the richer information environment that arose after the tick size reduction and to the ability of faster traders to exploit it. Following the tick size decrease, and owing importantly to the increase in liquidity consumption by HFTs, the role of the spot market in price discovery dropped relative to that of the futures markets. This points to the need for a balanced market ecology in financial markets where fast and slow traders coexist.

C1301: **Arbitrage and liquidity: Evidence from a panel of exchange traded funds**
Presenter: **David Rappoport**, Federal Reserve Board, United States
Co-authors: Tugkan Tuzun

Market liquidity is expected to facilitate arbitrage, which in turn should affect the liquidity of the assets traded by arbitrageurs. This relationship is studied by using a unique dataset of equity and bond ETFs compiled from big trade-level data. We find that liquidity is an important determinant of the efficacy of the ETF arbitrage. For less liquid bond ETFs, Granger-causality tests and impulse responses suggest that this relationship is stronger and more persistent, and liquidity spillovers are observed from portfolio constituents to ETF shares. The results inform the design of synthetic securities, especially when derived from less liquid instruments.

C1113: **When low-frequency measures really measures transaction costs**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States
Co-authors: Filip Zikes

Popular measures of transaction costs based on daily data with their high-frequency data-based counterparts are compared. We find that for U.S. equities and major foreign exchange rates, (i) the measures based on daily data are highly upward biased and imprecise; (ii) the bias is a function of volatility; and (iii) it is primarily volatility that drives the dynamics of these liquidity proxies both in the cross section as well as over time. We corroborate the results in carefully designed simulations and show that such distortions arise when the true transaction costs are small relative to volatility. Many financial assets exhibit this property, not only in the last two decades, but also in the previous century. We document that using low-frequency measures as liquidity proxies in standard asset pricing tests may produce sizable biases and spurious inferences about the pricing of aggregate volatility or liquidity risk.

CO626 Room G5 INFERENCE IN DATA-RICH ENVIRONMENTS: METHODS AND APPLICATIONS Chair: Andrew Butters

C1226: **Multi-sector business cycle accounting in a data-rich environment**
Presenter: **Scott Brave**, Federal Reserve Bank of Chicago, United States
Co-authors: Andrew Butters, David Kelley
Motivated by a multi-sector general equilibrium model with input-output linkages,we use a mixed-frequency structural dynamic factor model to decompose U.S. macroeconomic fluctuations into the contributions of four "wedges" commonly used in business cycle accounting: (i) an efficiency, (ii) a labor, (iii) an investment, and (iv) a government wedge. We then evaluate the extent to which shocks to these wedges identified from a mix of short- and long-run restrictions can explain the degree of cross-sectional co-movement in a panel of 500 real economic activity indicators at business cycle frequencies.

C1376: **Testing for changes in systematic correlation with approximate threshold group-factor models**
Presenter: **Mirco Rubin**, EDHEC - Nice, France
Co-authors: Daniele Massacci, Dario Ruzzi
The Approximate Linear Factor Model is extended in order to allow for different groups of individuals to be affected by both common and group-specific latent factors, with loadings switching over time as function of an exogenous variable. The model is inspired by the recent extensions of the classical approximate factor model allowing for either a) time-varying loadings with threshold-type regime switches or b) specific factors affecting only a finite number of groups. We propose inference procedures to detect i) the presence of switches in the loading structure, ii) the level of the exogenous variable determining the switches, iii) the number of factors which are across groups in each regime, as opposed to group-specific factors and iv) changes across regimes in the (average) systematic correlation among individuals in different groups due to the pervasive factors. Our measure of systematic correlation is related to the canonical correlations and the R-squared - via the loadings - of the estimated pervasive factors, and is used to test for changes in comovement between different groups across regimes.

C1398: **Now-casting macroeconomic trends and cycles**
Presenter: **Filippo Pellegrino**, LSE; Now-Casting Economics, United Kingdom
Co-authors: Lucrezia Reichlin, Giovanni Ricco, Thomas Hasenzagl
Building on recent developments in semi-structural econometric models, a new approach is proposed to nowcast key economic indicators and understand inflation dynamics. A mixed-frequency dataset is used that includes nominal, expectational, and real macroeconomic time series and timely survey data. These variables are processed in real-time, according to the official calendar of publications. As data becomes available, model estimates are updated to perform a timely trend-cycle decomposition and produce a new sequence of now-casts of the observed variables. The parameters are estimated via Bayesian methods with weakly informative priors. The models restrictions are informed by macroeconomic theory encompassing different hypotheses on the Phillips curve, the Okuns law, the permanent income hypothesis and the cyclical behaviour of oil prices.

CO410 Room Gordon TIME SERIES AND FORECASTING Chair: Robert Kunst

C0869: **Commodity market behavior in different states of the economy**
Presenter: **Ines Fortin**, Institute for Advanced Studies, Austria
Co-authors: Jesus Crespo Cuaresma, Jaroslava Hlouskova, Michael Obersteiner
Commodity prices have been identified as one of the main outstanding issues in the analysis of inflation. However, interpreting commodity price cycles and providing factor attribution is still a widely unsolved riddle. We plan to examine forecasting models for different commodity classes, where predictors include fundamental, macroeconomic and financial variables. The objective is to identify the role of market fundamentals such as inventories and financial market indicators in commodity price forecasts. To this end we will systematically compare a large battery of time series models including threshold models. In comparing the competing models, we use both traditional and more recent (profit-based) performance measures. The main objective is to find out whether the quality of commodity forecasts depends on the state of the economy and what variables are the key players in explaining different commodity classes in different states of the economy. We would like to see whether, for example, forecast models provide better predictions in calm than in turbulent times (periods of low/high volatility). Alternative states of the economy we would like to investigate are recessions/expansions, periods of high/low inflation, of high/low interest rates, and of different market sentiment.

C1667: **BVAR forecasts, survey information and structural change in the Euro area**
Presenter: **Florens Odendahl**, Banque de France, France
External information extracted from the European Central Banks Survey of Professional Forecasters is incorporated into the predictions of a Bayesian VAR, using entropic tilting and soft conditioning. Both methods significantly improve the plain BVAR point and density forecasts. Importantly, we do not restrict the forecasts at a specific quarterly horizon, but their possible paths over several horizons jointly, as the survey information comes in the form of one- and two-year-ahead expectations. Besides improving the accuracy of the variable that we target, the spillover effects to other-than-targeted variables are relevant in size and statistically significant. We document that the baseline BVAR exhibits an upward bias for GDP growth after the financial crisis and our results provide evidence that survey forecasts can help mitigate the effects of structural breaks on the forecasting performance of a popular macroeconometric model.

C0906: **Forecasting agricultural product and energy prices: A simulation-based model selection approach**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Adusei Jumah
The aim is twofold. First, we study whether and to what degree the dynamic interaction between commodity prices and energy prices can be exploited for forecasting. Second, we present informative examples for the simulation-based forecast-model selection procedure. Apart from prediction by competing specifications to be selected from a small choice set, we also explore forecast combinations based on Bates-Granger weights constructed from a continuum in the same framework. The simulation-based method explicitly permits letting the forecast model choice depend on the intended time horizon of the forecast. With regard to classical Granger causality, the evidence supports a causal direction from food prices to fuel prices, without feedback and somewhat in contrast to our expectations. This causal link, however, only benefits forecasting accuracy at relatively large sample sizes. Similarly, clear evidence on considerable seasonal patterns cannot be fused to a seasonal time-series model that outperforms non-seasonal rivals. The simulation experiments generally favor the handling of all price series in first differences. Ultimately, the forecast combination experiments indicate a window of opportunity at a specific horizon, whereas pure strategies dominate at smaller and larger horizons.

CO757 Room Woburn TOPICS IN MACRO AND FINANCE Chair: Alessia Paccagnini

C0919: **Impact of bail-in on banks' bond yields and market discipline**
Presenter: **Raffaele Giuliana**, City University London and Central Bank of Ireland, United Kingdom
The EU statutory bail-in regime attempts to promote market discipline and mitigate the too-big-to-fail problem by limiting governments' support for equity-holders and unsecured debt-holders of failing banks. The purpose is to analyze staggered events relating to the legislative process of the bail in and its impositions on failing banks. We test if these events modified bail-in expectations among bond-holders. Difference-in-differences tests suggest that the events indicating an increased commitment to bail-in increased the difference in yield between unsecured (i.e., bailinable)

and secured (i.e., non-bailinable) bonds. These results are not driven by the possible generalized instability associated with bail-ins. In addition, triple-differencing framework shows that bail-in was more effective for larger banks and that it improved market discipline.

C0946: **A credit-based theory of the currency risk premium**
Presenter: **Pasquale Della Corte**, Imperial College London, United Kingdom
Co-authors: Alexandre Jeanneret, Ella Patelli

A novel component for exchange rate predictability is uncovered. Our theory shows that currency returns compensate investors for the expected currency depreciation in the case of a severe but rare credit event. We compute this risk compensation the credit-implied risk premium (CRP) by exploiting the price difference between sovereign credit default swaps denominated in different currencies. Using data for 16 Eurozone countries over the period 2010-17, we find that CRP positively forecasts the euro-dollar exchange rate return between one-week and six-month horizon, both in-sample and out-of-sample. We also show that currency trading strategies that exploit the informative content of CRP generate substantial out-of-sample economic value.

C0982: **Portfolio choice under uncertainty**
Presenter: **Isabella Blengini**, Ecole hoteliere de Lausanne (EHL), Switzerland
Co-authors: Alessia Paccagnini

The factors that affect agents' portfolio choices when there is an increase in world uncertainty are analyzed. We solve a DSGE model with uncertainty using the method developed previously. Thanks to our macroeconomic setting, we can endogenously determine asset returns and clarify their relationship with the macroeconomic fundamentals. In our two-country DSGE model we assume that there is trade in both goods and financial assets. The international asset portfolio includes two types of securities: stocks and bonds. Each country issues one government bond and one equity, denominated in local goods. There are three sources of shocks: one preference shock that is common to the two economies, two endowment shocks and two government spending shocks. We proxy the increase in uncertainty with the introduction of uncertainty shocks, i.e., we allow the variances of the shocks to be time-varying. Investors choose their portfolio with one main goal in mind: They want to smooth their consumption. When the uncertainty shocks hit, the way in which real variables co-vary with asset returns changes. As a consequence, agents need to re-adjust their portfolios until when the shock disappears. That is why we observe portfolio dynamics. Our main findings suggest that the response of the portfolio to an increase in uncertainty crucially depends on the source of uncertainty.

CO620 Room Chancellor's Hall ADVANCES IN NONPARAMETRIC AND SEMIPARAMETRIC ECONOMETRICS Chair: Francesco Bravo

C0714: **A doubly corrected robust variance estimator for linear GMM**
Presenter: **Byunghoon Kang**, Lancaster University, United Kingdom

A new finite sample corrected variance estimator for the linear generalized method of moments (GMM) is proposed including the one-step, two-step, and iterated estimators. The formula additionally corrects for the over-identification bias on top of the commonly used finite sample Windmeijer correction, which corrects for the bias from estimating the efficient weight matrix, so is doubly corrected. The over identification bias arises from the fact that the over-identified sample moment condition is nonzero in finite sample while it converges in probability to zero under correct specification. The order of the over-identification bias equals the order of the sample moment condition. Thus, our double correction is higher-order under correct specification. However, our double correction becomes first-order under misspecification because the sample moment condition does not converge in probability to zero. This implies that the conventional variance estimator and the Windmeijer correction are inconsistent, while our doubly corrected variance estimator is consistent even when the moment condition model is misspecified. That is, the proposed formula provides a convenient way to obtain improved inference under correct specification and robustness against misspecification at the same time.

C0735: **Jackknife, small bandwidth and high-dimensional asymptotics**
Presenter: **Taisuke Otsu**, London School of Economics, United Kingdom
Co-authors: Yukitoshi Matsushita

Light is shed on problems of statistical inference under alternative or nonstandard asymptotic frameworks from the perspective of jackknife empirical likelihood (JEL). Examples include small bandwidth asymptotics for semiparametric inference, many covariates asymptotics for regression models, and many-weak instruments asymptotics for instrumental variable regression. We first establish Wilks' theorem for the JEL statistic on a general semiparametric inference problem under the conventional asymptotics. We then show that the JEL statistics lose asymptotic pivotalness under the above nonstandard asymptotic frameworks, and argue that these phenomena are understood as emergence bias of the jackknife variance estimator in the first order. Finally, we propose a modification of JEL to recover asymptotic pivotalness under both the conventional and nonstandard asymptotics. Our modification works for all above examples and provides a unified framework to investigate nonstandard asymptotic problems.

C0708: **Misspecified semiparametric models selection**
Presenter: **Francesco Bravo**, University of York, United Kingdom

Tests are proposed for comparing two possibly misspecified semiparametric moment conditions models with weakly dependent data. The asymptotic distributions of the resulting test statistics are not standard but can be consistently estimated by a particular weighted bootstrap procedure. The results are illustrated with a simulation study and an empirical application.

CO418 Room Jessel NEW DEVELOPMENTS IN FINANCIAL TIME SERIES Chair: Richard Gerlach

C1457: **Asymptotic properties of mildly explosive processes with locally stationary disturbance**
Presenter: **Junichi Hirukawa**, Niigata University, Japan
Co-authors: Sangyeol Lee

The aim is to derive the limiting distribution of the least squares estimator (LSE) and the localized LSE for mildly explosive autoregressive models with locally stationary disturbance and verify that it is Cauchy as in the iid case. We also investigate the limiting distribution of two types of Dickey-Fuller unit root tests, designed for detecting a bubble period in economic time series data, and show that these tests are consistent. To evaluate the methods, we conduct a simulation study and carry out a data analysis using time series data on bitcoin prices.

C1662: **Estimating multiple quantiles via a reparametrized stochastic gradient algorithm**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan

A method is proposed for simultaneously estimating several quantile regressions with the same covariates. The method reparametrizes predictors by allocating observed covariates to different containers according to whether they are positively valued or negatively valued. The resulting regressions will consist two sets of covariates that are positively-valued and negatively-valued, respectively. The reparametrization allows us to control the order of regression coefficients without considering the sign of the corresponding covariates. The method adopts an iterative scheme based on the subgradient algorithm for obtaining a solution to the estimation problem. The method is then applied to multiple quantile estimation for time series data.

C1586: **A resampling method for dynamic quantile models of asset returns**
Presenter: **Richard Luger**, Laval University, Canada

Suppose the joint distribution of daily returns is symmetric and a consistent point estimator is available for the parameters of a dynamic quantile

model of the asset's multi-day returns. The considered class of dynamic quantile models includes linear and non-linear autoregressive specifications. In this setting, a simple and general resampling method is proposed to obtain the distribution of parameter estimates, which may be constrained to avoid the crossing problem when several quantile levels are fitted. With large sample sizes, the resampling distribution allows the construction of simultaneous confidence intervals for continuous functions of the model parameters. The usefulness of this non-parametric inference procedure is illustrated by means of a simulation study and with an empirical application featuring a conditional autoregressive value-at-risk (CAViaR) model for daily returns and a quantile autoregression (QAR) model for longer horizons.

CG019 Room Montague CONTRIBUTIONS IN COINTEGRATION Chair: Simon Clinet

C1845: **Trend IV estimation and inference in cointegrating regressions: Basic properties and some extensions**
Presenter: **Julio Angel Afonso-Rodriguez**, University of the Balearic Islands, Spain

With only very few recent exceptions, consistent and asymptotically efficient estimation and inference of cointegrated regression models under general conditions, namely, with serially correlated stationary error terms and endogenous integrated regressors, largely depend on the choice of several tuning parameters and usually requires the preliminary estimation by OLS. One of such recent approaches, called Trend Instrumental Variable (TIV) estimation, based on an apparently spurious trend regression with deterministic trend instruments derived from the Karhunen-Love representation of a Brownian Motion and applied to an augmented version of the basic cointegrating regression model with no loss of degrees of freedom in the estimation, is reviewed in detail and extended to some relevant cases. First, in the most restricted case without deterministic components, we study the treatment of different assumptions on the initial condition for exactly integrated regressors, and on the other hand the effects of including a subset of stationary or cointegrated regressors (subcointegration). Second, we analyze the extension to the case of inclusion of deterministic components in the regression when the regressors are deterministically trending integrated. Third, we consider the case of a cointegrating regression with threshold effects. Finally, in all these cases we study the construction and properties of several statistics based on TIV residuals to test for cointegration.

C0267: **Return predictability from upside and downside variance premia: A fractionally co-integrated analysis**
Presenter: **Marwan Izzeldin**, Lancaster University Management School, United Kingdom
Co-authors: Xingzhi Yao

The realized and implied variances are decomposed into upside and downside semi-variances to better capture the variation in market compensation during good and bad uncertainties. We show that the fractional co-integration between implied and realized variances is driven by the downside components whereas that relationship is not observed for the upside components. Moreover, return predictability inherent in the downside variance risk premium (VRP) dominates that of the total VRP. This finding is further verified using a fractionally co-integrated VAR framework.

C1508: **Fractional trends in unobserved components models**
Presenter: **Tobias Hartl**, University of Regensburg, Germany
Co-authors: Rolf Tschernig, Enzo Weber

A generalization of permanent-transitory decompositions is developed that avoids prior assumptions about the long-run dynamic characteristics by modelling the permanent component as a fractionally integrated process and incorporating a fractional lag operator into the autoregressive polynomial of the cyclical component. The model neither requires stationarity nor orthogonal permanent and transitory shocks and can be cast in state-space form. In a multivariate setup, fractional trends may exhibit different integration orders, but depend on the same stochastic shocks, leading to a cointegrated system with different integration orders. We show that our fractional UC model is able to estimate a smooth trend together with a cycle hitting all NBER recessions for US real output. In a multivariate setup, we provide evidence that income and inflation are driven by the same long-run shocks, although their trend components are of different persistence.

CG021 Room Court CONTRIBUTIONS ON COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Stephen Pollock

C2018: **A non-linear filter for output gaps and bear markets**
Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy

The focus is on the non linear filter that arises from subtracting the running maxi-mum of the last q observations of a continuous stochastic process from the current value. The filter has applications in both macroeconomics and finance, providing a measure of the depth of a recession and of a bear market, respectively. In image processing the maximum filter is known as a dilation filter. We show that the maximum filter defines a Markov Chain with $q+1$ states that is homogeneous and has a stationary distribution for difference stationary processes and we derive the ergodic and transition probabilities as a function of the mean and the autocovariance function of the first differences of the process. Hence, we show that the depth variable is a covariance stationary random process and derive its moments. A new test of the efficient market hypothesis is proposed that looks at the symmetry of the ergodic probability distribution.

C1666: **Enhanced methods of seasonal adjustment**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

The effect of the conventional model-based methods of seasonal adjustment is to nullify the elements of the data that reside at the seasonal frequencies and to attenuate the elements at the adjacent frequencies. It may be desirable to nullify some of the adjacent elements instead of merely attenuating them. For this purpose, two alternative procedures are presented that have been implemented in a computer program. In the first procedure, the seasonal-adjustment filter is augmented by additional poles and zeros that are targeted at the adjacent frequencies. In the second procedure, a Fourier transform is deployed to reveal the elements of the data at all the frequencies. This allows the elements in the vicinities of the seasonal frequencies to be eliminated or attenuated at will. In spite of the success of these procedures, the question is raised of whether the estimated trend-cycle trajectory can serve in place of the seasonally adjusted data.

C2034: **Forecasting with news sentiment: Evidence from UK newspapers**
Presenter: **Dooruj Rambaccussing**, University of Dundee, United Kingdom

The performance of newspapers for large scale macro-forecasting in the United Kingdom is investigated by introducing new time series of economic sentiments- Economic Policy Sentiment and Broad Economic Sentiment. Information contained in pre-selected newspaper articles are extracted using textual techniques (Nave Bayes Classier and Support Vector Machines.) The findings of the paper are fourfold. Firstly, within the set of soft variables, Superior Model Confidence Sets and Encompassing Tests show that newspapers can be used in concert with other survey data measures. Secondly, newspapers are useful for short-term forecasting in times of economic uncertainty. Thirdly, newspapers, at an individual level, differ in forecasting performance across the variables being forecast. Finally, newspaper articles containing policy relevant terms are better forecasters than broad economic sentiments.

Authors Index

