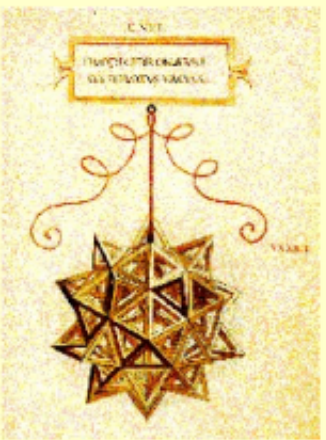
Joint Meeting

GfKl – CLADAG 2010

8-10 September 2010

Firenze (Italy)



Book of Abstracts

Joint Meeting

GfKl - CLADAG 2010

8-10 September 2010
Firenze (Italy)

Book of Abstracts

GfKl-CLADAG 2010 COMMITTEES

Scientific Program Committee
Antonio Giusti (co-chair)
Gunter Ritter (co-chair)
Daniel Baier
Reinhold Decker
Filippo Domma
Luigi Fabbris
Andrea Giommi
Christian Hennig
Carlo Lauro
Berthold Lausen
Hermann Locarek-Junge
Isabella Morlini
Lars Schmidt-Thieme
Gabriele Soffritti
Alfred Ultsch
Rosanna Verde
Donatella Vicari
Claus Weihs

Local Organizing Committee
Andrea Giommi (chair)
Bruno Bertaccini
Matilde Bini
Chiara Bocci
Antonio Giusti
Anna Gottard
Leonardo Grilli
Riccardo Innocenti
Alessandra Mattei
Alessandra Petrucci
Carla Rampichini
Emilia Rocco

GfKl-CLADAG 2010 Conference Agenda

	Wednesday, September 8, 2010
9:00 9:30	**Opening** **Ceremony (0.18)**
9:30 10:45	**Contributed** **01 - Achievement and research (0.04)** **02 - Applications in economics (0.05)** **03 - Dimension reduction (0.18)** **04 - Environmental data analysis and classification (0.07)** **05 - Applications in demography (0.06)**
10:45 11:15	**Coffee break**
11:15 12:45	**Specialized** **A - Challenges and advances in statistical spatial data analysis (0.04)** **B - Robust statistical methods for data analysis: methodological innovations and applications (0.07)** **C - Web data mining (0.18)**
12:45 14:15	**Lunch**
14:15 15:30	**Contributed** **06 - Handwriting and web analysis (0.05)** **07- Applications to sociology and market (0.06)** **08 - Discrete data (0.04)** **09 - Classification methods (0.18)** **10 - Time series and spatial analysis (0.07)**
15:30 16:30	**First plenary** M. Weber, S. Röblitz, F. Haack **Adaptive spectral clustering in molecular simulation (0.18)**
16:30 17:00	**Coffee break**
17:00 18:30	**Specialized** **D - Opinion mining and preference analysis (0.04)** **E - Recent developments in recursive partitioning methods (0.07)** **F - Statistical signal analysis (0.18)**
19:30	***Welcome***

	Thursday, September 9, 2010
8:30 9:45	**Contributed** **11 - Advances in latent class modelling (0.18)** **12 - Clustering and dissimilarities (0.04)** **13 - Customer satisfaction and conjoint analysis (0.05)** **14 - Recent trends in classification (0.07)** **15 - Regression and factor analysis (0.06)**
9:45 10:45	**Second plenary** A.-L. Boulesteix **Critical issues and developments in microarray-based prediction (0.18)**
10:45 11:15	**Coffee break**
11:15 12:45	**Specialized** **G - Classification in systems biology I (0.18)** **H - Data stream mining (0.04)** **I - Data visualisation (0.07)**
12:45 14:15	**Lunch**
14:15 15:30	**Contributed** **16 - Biomedical applications (0.18)** **17 - Developments in archetypal analysis (0.04)** **18 - Model selection (0.07)** **19 - Proximity data and hierarchies (0.06)** **20 - Professional profiles (0.05)**
15:30 16:30	**Third plenary** V. Esposito Vinzi, M. Zargoush **Knowledge extraction by investigating model uncertainty through predictive path modeling and probabilistic networks (0.18)**
16:30 17:00	**Coffee break**
17:00 18:30	**Specialized** **J - Classification in systems biology II (0.18)** **K - Social networks and classification (0.04)** **L - Statistical matching: theory and applications to data mining and official statistics (0.07)**
20:00	***Gala dinner at "Biblioteca delle Oblate"***

	Friday, September 10, 2010
8:30 9:45	**Contributed** **21 - Correspondence analysis and related methods I (0.04)** **22 - Estimation problems (0.05)** **23 - Issues in classification and clustering (0.18)** **24 - Multivariate analysis for relational data (0.07)** **25 - Rankings and preferences (0.06)**
9:45 10:45	**Fourth plenary** R. Rocci **Mixing mixtures of Gaussians (0.18)**
10:45 11:15	**Coffee break**
11:15 12:30	**Contributed** **26 - Correspondence analysis and related methods II (0.04)** **27 - Methodology and applications of latent class and mixture models (0.18)** **28 - Markov and graphical modeling (0.07)** **29 - Variable structures (0.05)** **30 - Risk analysis (0.06)** **31 - Miscellanea (0.15)**
12:30 13:00	**Closing Ceremony (0.18)**

	Duration	Presentations for session	Time for each presentation	Time for president and floor discussion
Plenary Sessions	$1^h00'$	1	40’	20’
Specialized Sessions	$1^h30'$	3 + discussant	3x20’+15’	15’
Contributed Sessions	$1^h15'$	3	20’	15’

Wednesday, September 8, 2010

9:00-9:30 Opening ceremony

Room: 0.18

A. Tesi (Rector of the University of Firenze), F. Giunta (Dean of the Faculty of Economics), S. Salvini (Head of the Department of Statistics "G. Parenti"), C. Weihs (President of GfKl), A. Cerioli (President of CLADAG)

9:30-10:45 Contributed sessions

01 Achievement and research
Chair: C. Rampichini Room: 0.04

F. De Battisti, S. Salini
Bibliometric indicators for statisticians: critical assessment in the italian context

I. Sulis, M. Porcu
Measuring the effect of cultural capital on students' achievement

E. Zavarrone
Formative measurement model for academic reputation

02 Applications in economics
Chair: H. Locarek-Junge Room: 0.05

P. Chirico
A regression clustering method for the prediction of the pro capita disposal income in municipalities

F. Cipollini, C. Ferretti, P. Ganugi, M. Mezzanzanica
Discrete and continuous time mover-stayer model for labour market in a small northern Italian area

A. A. Romano, G. Scandurra
Electricity consumption and gross domestic product in the Italian regions

03 Dimension reduction
Chair: A. Cerioli Room: 0.18

S. Borra, A. Di Ciaccio
Variable selection in a predictive approach

K. Luebke, C. Weihs
Adaptive linear dimension reduction in a classification setting

R. Rocci, S. A. Gattone
Dimensional reduction and clustering of functional data

04 Environmental data analysis and classification
Chair: A. Petrucci Room: 0.07

E. Di Giuseppe, G. Jona Lasinio, M. Pasqui, S. Esposito
Functional clustering of temperature and precipitation data for Italian climate zones determination

E. Nissi, A. L. Sarra, S. Palermi, G. De Luca
The application of M-function analysis to the geographical distribution of earthquake sequence

A. Plaia, M. Ruggieri, F. Di Salvo, G. Agro
From a multivariate spatio-temporal array to a multipollutant - multisite air quality index

05 Applications in demography
Chair: S. Salvini Room: 0.06

S. Bozza, M. Di Bacco, R. Bigazzi, S. De Iasio
Analysis of the individual variability of sex ratio with hierarchical models

S. Korenjak-Cerne, N. Kejzar, V. Batagelj
Clustering of population pyramids presented as histogram symbolic data

D. Vignoli, A. Mattei, A. Gottard
Modeling fertility and education in Italy: time-variant or invariant unobserved heterogeneity component?

10:45-11:15 Coffee break

11:15-12:45 Specialized sessions

A Challenges and advances in statistical spatial data analysis
Chair: A. Petrucci Discussant: A. Ultsch Room: 0.04

M. Behnisch, A. Ultsch
Urban knowledge discovery – Swiss population development by 15 decades

B. Cafarelli, A. Pollice, G. Jona Lasinio
Agronomic field grain property maps by geoadditive models: a comparison of different spatial correlation structures

M. Pratesi, S. Marchetti, C. Giusti, N. Salvati
Spatial model in small area estimation: an M-quantile approach

B Robust statistical methods for data analysis: methodological innovations and applications
Chair: L. Grossi Discussant: G. Ritter Room: 0.07

G. Cavaliere, I. Georgiev
Exploiting infinite variance through dummy variables in an AR model

D. Ferrari, D. La Vecchia
On robust estimation via pseudo-additive information

R. Hable, A. Christmann
Robustness versus consistency in ill-posed classification and regression problems

C Web data mining
Chair: A. Geyer-Schulz Discussant: F. Palumbo Room: 0.18

F. Camillo, F. Neri
Monitoring the web sentiment, the Italian Prime Minister's case

W. Gaul
Web page importance ranking

Thai-Nghe Nguyen, Z. Gantner, L. Schmidt-Thieme
Evaluation metric for learning from imbalanced data based on asymmetric beta distribution

12:45-14:15 Lunch

14:15-15:30 Contributed sessions

06 Handwriting and web analysis
Chair: W. Gaul Room: 0.05

S. Bozza, F. Taroni, R. Marquis, M. Schmittbuhl
The evaluation of handwriting evidence: multi-level models for determining authorship

R. Kenett, S. Salini
Relative linkage disequilibrium in tracking web search patterns

K. Silachan, P. Tantatsanawong, C. Lursinsap
Bayesian classification tree for statistical user web-URL categories navigation pattern model

07 Applications to sociology and market
Chair: R. Decker Room: 0.06

M. De Castris, G. Pellegrini
Integrating the spatial dimension into propensity score matching to evaluate regional impact of capital subsidies

C. Paccagnella, R. Varriale
Asset ownership of the elderly across Europe: a multilevel latent class analysis to segment country and households.

M. T. Santoro, S. Staffieri
Restructuring and innovations on the survey "capacity of collective tourist accommodation" and their impact on the process quality

08 Discrete data
Chair: C. Rampichini Room: 0.04

P. A. Ferrari, A. Barbiero
Generating ordinal data

P. Giudici, E. Raffinetti
Gini measure: its decomposition proposal in the discrete case

A. Mazza, A. Punzo
Adaptive discrete beta kernel graduation of demographic data

09 Classification methods
Chair: F. Palumbo Room: 0.18

H. Hruschka
Restricted Boltzmann machines for market basket analysis

A. Irpino, M. R. Guarracino, R. Verde
Classification of chunked data using proximal vector machines and singular value decomposition

N. Louw
Robust kernel Fisher discriminant analysis with weighted kernels

10 Time series and spatial analysis
Chair: H. Locarek-Junge Room: 0.07

C. Drago, G. Scepi
Clustering and forecasting beanplot time series

M. Mucciardi, P. Bertuccelli
Local analysis of spatial relationships: a comparison of GWR and OLS method

L. Spezia
Classification of spatio-temporal series and hidden Markov models

15:30-16:30 First plenary session

Chair: G. Ritter Room: 0.18

M. Weber, S. Röblitz, F. Haack
Adaptive spectral clustering in molecular simulation

16:30-17:00 Coffee break

17:00-18:30 Specialized sessions

D Opinion mining and preference analysis
Chair: R. Decker Discussant: P. Giudici Room: 0.04

G. Giordano
On the use of multivariate multiple regression models in the elicitation of consumer preferences

M. Iannario, D. Piccolo
A model-based approach for qualitative assessment in opinion mining

D. Schindler, L. Lüpke
Preference analysis for durable goods – Some remarks on data consistency and empirical results

E Recent developments in recursive partitioning methods
Chair: M. Pillati Discussant: B. Lausen Room: 0.07

H.-H. Bock
An iterative fuzzy and time-dynamic clustering approach for time series

C. Conversano
Evaluating the performance of different sets of classifiers in multiclass learning

R. Siciliano
Recursive partitioning of complex data for classification and regression trees

F Statistical signal analysis
Chair: C. Weihs Discussant: R. Verde Room: 0.18

M. Eichhoff
Musical instrument detection based on extended feature analysis

K. Friedrichs, C. Weihs
Auralization of auditory models

E. Romano, A. Irpino
Spatially constrained curve clustering: a hierarchical approach to signals analysis

From 19:30 Welcome

Thursday, September 9, 2010

8:30-9:45 Contributed sessions

11 Advances in latent class modelling
Chair: L. Grilli Room: 0.18

F. Bartolucci, F. Pennoni, L. Pieroni
A latent class version of the inverse probability-to-treatment weighted estimator for dynamic causal effects

B. Grün, K. Hornik
Finite mixture modeling of censored longitudinal data

R. Varriale, J. K. Vermunt
Determining the number of components in multilevel mixture (factor) models

12 Clustering and dissimilarities
Chair: G. Soffritti Room: 0.04

L. Cutillo, A. Carissimo, D. di Bernardo
Optimal length choice in top k ordered lists aggregation

B. Fichet
Recognition of ultrametrics and tree-metrics in optimal time

I. Morlini, S. Zani
A dissimilarity measure between two hierarchical clusterings

13 Customer satisfaction and conjoint analysis
Chair: L. Fabbris Room: 0.05

G. Boari, G. Cantaluppi
Unveiling non-linear relationships between perceived satisfaction and quality

A. De Luca
Ordinal logistic regression for the estimate of the response functions in the conjoint analysis

P. Kurz, A. Sikorski
Individual Self Balancing Conjoint (ISBC): an adaptive design technique for choice based conjoint

14 Recent trends in classification
Chair: A. Hardy Room: 0.07

A. Balzanella, R. Verde, Y. Lechevallier
Clustering highly evolving data streams

G. Cabanes, Y. Bennani
Extending SOM with efficient estimation of the number of clusters

D. C. Porumbel, J. K. Hao, P. Kuntz
A classification approach for structure discovery in search spaces of combinatorial optimization problems

15 Regression and factor analysis
Chair: R. Siciliano Room: 0.06

R. Calabrese, S. A. Osmetti
Generalized extreme value regression in rare events

G. Giordano, M. Aria
Ensemble procedures for more accurate regression trees with moderating effects

J. Ohrvik, G. Schoier
On the use of bootstrap in factor analysis

9:45-10:45 Second plenary session

Chair: M. Vichi Room: 0.18

A.-L. Boulesteix
Critical issues and developments in microarray-based prediction

10:45-11:15 Coffee break

11:15-12:45 Specialized sessions

G Classification in systems biology I
Chair: B. Lausen Discussant: M. Marchi Room: 0.18

A. Benner
Penalized likelihood approaches for high-dimensional model selection

E. Dreassi
A shared components model to detect uncommon risk factors in disease mapping

M. Metodiev
Making sense of large-scale proteomics datasets: myths, facts, and challenges associated with the analysis and statistical evaluation of protein ID and abundance data

H Data stream mining
Chair: A. Irpino Discussant: C. Weihs Room: 0.04

A. Balzanella, R. Verde
Clustering highly evolving data streams

F. Clerot, P. Gouzien
Information-based data stream summary

K. Tschumitschew, F. Klawonn
Tests for change detection based on incremental quantile estimation

I Data visualization
Chair: R. Rocci Discussant: C. Hennig Room: 0.07

A. Gribov, A. Unwin
Visualization of clustering comparisons with confusion matrices in Seurat

J. Mucha, H.-G. Barte
Visualisation of cluster analysis results

L. Scrucca
Some recent advances on dimension reduction for high-dimensional data

12:45-14:15 Lunch

14:15-15:30 Contributed sessions

16 Biomedical applications
Chair: T. R. Lee Room: 0.18

D. De March, I. Poli
Evolutionary neural networks to design synthetic proteins

M. Ouedraogo, F. Lecerf, S. Lê
Understanding co-expression of co-located genes using a PCA approach

I. Rocchetti
Modeling delay to diagnosis for amiotrophic lateral sclerosis: under reporting and incidence estimates

17 Developments in archetypal analysis
Chair: F. Palumbo Room: 0.18

S. Corsaro, M. Marino
Archetypal analysis for interval valued data

M. R. D'Esposito, F. Palumbo, G. Ragozini
Archetypal analysis for prototype identification

M. Eugster, F. Leisch
New features for archetypal analysis in R

18 Model selection
Chair: H.-H. Bock Room: 0.07

B. Bertaccini, F. Polverini
Automatic detection of outliers in linear regression models: the forward search approach

D. Facchinetti, S. A. Osmetti
Estimating the gap of the forward search via censored sampling

R. Savona, M. Vezzoli
Assessing model accuracy using a two-dimensional loss function

19 Proximity data and hierarchies
Chair: I. Morlini Room: 0.06

L. Bocci
Multi-objective genetic algorithm based clustering for dissimilarity data

M. Cadoret, S. Lê, J. Pagès
A new approach for analyzing a set of hierarchies

A. Geyer-Schulz, M. Ovelgőnne
Extensions of modularity clustering

20 Professional profiles
Chair: L. Grilli Room: 0.05

M. Civardi, F. Crippa, V. Bagnardi
University graduate's job hunting: what helps to get the good one?

L. Fabbris, G. Boccuzzo
The detection of criticalities of graduates' jobs through importance-performance analysis

C. Martini
Statistical methods to describe professional profiles through competences and activities

15:30-16:30 Third plenary session

Chair: A. Giusti Room: 0.18

V. Esposito Vinzi, M. Zargoush
Knowledge extraction by investigating model uncertainty through predictive path modeling and probabilistic networks

16:30-17:00 Coffee break

17:00-18:30 Specialized sessions

J Classification in systems biology II
Chair: B. Lausen Discussant: A. Biggeri Room: 0.18

M. Göker
Phylogenomics as a standard technique for microbial taxonomy

H. Kestler
Boolean networks for modeling gene regulation

F. Stefanini
Graphical models for eliciting structural information

K Social networks and classification
Chair: M. R. D'Esposito Discussant: A. Geyer-Schulz Room: 0.04

U. Brandes, J. Lerner, M. J. Lubbers, C. McCarty, J. Luis Molina, U. Nagel
Classification of personal networks using latent roles

A. Ferligoj
Developments in blockmodeling

G. Giordano, M. P. Vitale
Clustering social actors by using auxiliary information on relational data

L Statistical matching: theory and applications to data mining and official statistics
Chair: N. Torelli Discussant: V. Esposito Vinzi Room: 0.07

M. D'Orazio, M. Di Zio, M. Scanu
Matching two different topics: ecological inference and data fusion

F. Meinfelder, S. Rässler
Displaying uncertainty in data fusion by imputation

P. Van der Putten
Data fusion for direct marketing

From 20:00 Gala dinner at "Biblioteca delle Oblate"

Friday, September 10, 2010

8:30-9:45 Contributed sessions

21 Correspondence analysis and related methods I
Chair: J. Blasius Room: 0.04

S. Camiz, G. Coelho Gomes
Joint correspondence analysis vs. multiple correspondence analysis: a solution to an undetected problem

M. Greenacre
Measuring subcompositional incoherence in contingency tables and compositional data

A. Langovaya, H. Chouikha, S. Kuhnt
Correspondence analysis in the case of outliers

22 Estimation problems
Chair: A. Mattei Room: 0.05

G. Mellace, R. Rocci
Principal stratification in sample selection problems with non normal error terms

M. Montinaro, I. Sciascia
Some considerations on two-stage calibration estimators

N. Solaro
Automatic strategies of analysis for handling structurally missing occasions

23 Issues in classification and clustering
Chair: D. Vicari Room: 0.18

J. Dias
Model-based clustering of multistate data with latent change. An application with DHS data

J. Heikkonen, D. Perrotta, M. Riani, F. Torti
Issues on clustering and data gridding

V. A. Tutore, V. Cozza, A. D'Ambrosio
Tree partitioning criteria across objects and predictors for data with a double stratification

24 Multivariate analysis for relational data
Chair: G. Giordano Room: 0.07

V. Batagelj
Cluster analysis of multivariate relational data

D. De Stefano, G. Ragozini
Multiple correspondence analysis for relational data

J. Lerner
Analysis of multivariate event networks

25 Rankings and preferences
Chair: D. Piccolo Room: 0.06

B. Arpino, R. Varriale
A Monte-Carlo study to evaluate value-added models for institutions' rankings

P. Cerchiello, P. Giudici
Ordinal models to assess media reputation

L. Deldossi, R. Paroli
Inference on the CUB model: a MCMC approach

9:45-10:45 Fourth plenary session

Chair: C. Hennig Room: 0.18

R. Rocci
Mixing mixtures of Gaussians

10:45-11:15 Coffee Break

11:15-12:30 Contributed Sessions

26 Correspondence analysis and related methods II
Chair: M. Greenacre Room: 0.04

S. Balbi, M. Misuraca, E. Zavarrone
Comparing mental maps: Obama vs. McCain

J. Blasius
Assessing the response quality in ordered categorical data

A. Iodice D'Enza, F. Palumbo
Adaptive factorial clustering for binary data

27 Methodology and applications of latent class and mixture models
Chair: C. Rampichini Room: 0.18

R. Arboretti Giancristofaro, S. Bonnini, E. Grossule, S. Ragazzi, L. Salmaso
Statistical cognitive survey on Passito wine in Veneto region (Italy) from the consumer's point of view

S. Bianconcini, S. Cagnone, P. Monari
Covariate effects in multivariate latent growth models for the analysis of undergraduated student performances

M. Matteucci, S. Mignani, B. P. Veldkamp
On using item features to estimate parameters in IRT models

28 Markov and graphical modelling
Chair: A. Gottard Room: 0.07

M. Costa, L. De Angelis
Model selection in latent Markov models: a simulation study

S. Pandolfi, F. Bartolucci, A. Farcomeni
Bayesian analysis of longitudinal categorical data via latent Markov models

E. Stanghellini, B. Vantaggi
On the identification of discrete graphical models with hidden nodes

29 Variable structures
Chair: F. Palumbo Room: 0.05

S. Camiz, J.-J. Denimal
Hierarchical factorial classification of variables: methods and applications

C. Davino, R. Romano
Sensitivity analysis of composite indicators through mixed model anova

K. Sahmer
A model for the clustering of variables taking into account external data

30 Risk analysis
Chair: M. Riani Room: 0.06

S. Facchinetti, P. Giudici, S. A. Osmetti
The distribution of the stochastic dominance index for risk measurement

S. Figini, P. Giudici, P. Uberti
Concentration measures for risk analysis

S. Figini, L. Grossi
Robust estimation and prediction for credit risk models

31 Miscellanea
Chair: A. Giommi Room: 0.15

F. Benassi, C. Bocci, A. Petrucci
Spatial clustering for local analysis

T. R. Lee, D. G. Jeon, E. Diday
Symbolic tree for prognosis of localized osteosarcoma patient

12:30-13:00 Closing ceremony

Room: 0.18

G. Ritter and A. Giusti (Co-Chairs of the S.P.C.), A. Giommi (Chair of L.O.C.)

GfKl-CLADAG 2010 Conference venue

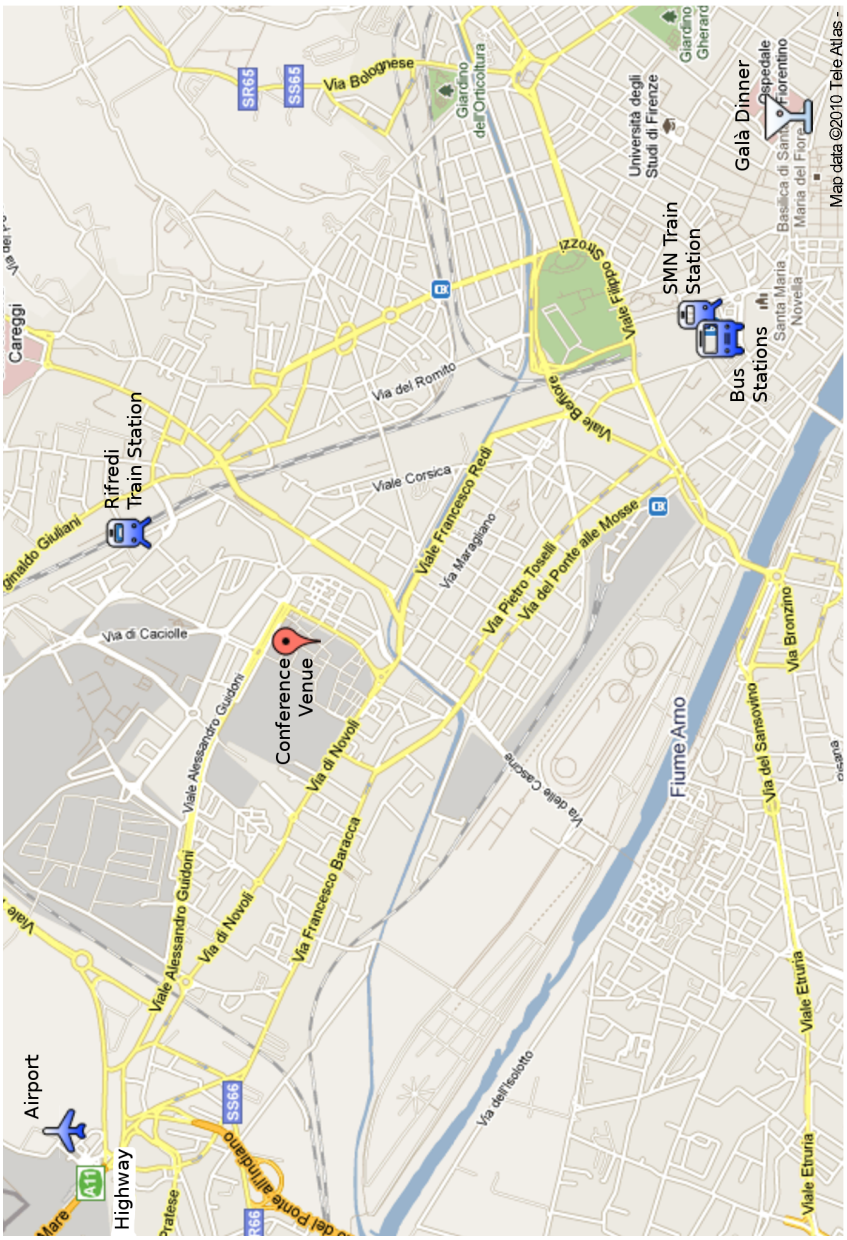


Fig. 1 GfKl-CLADAG 2010 Conference venue

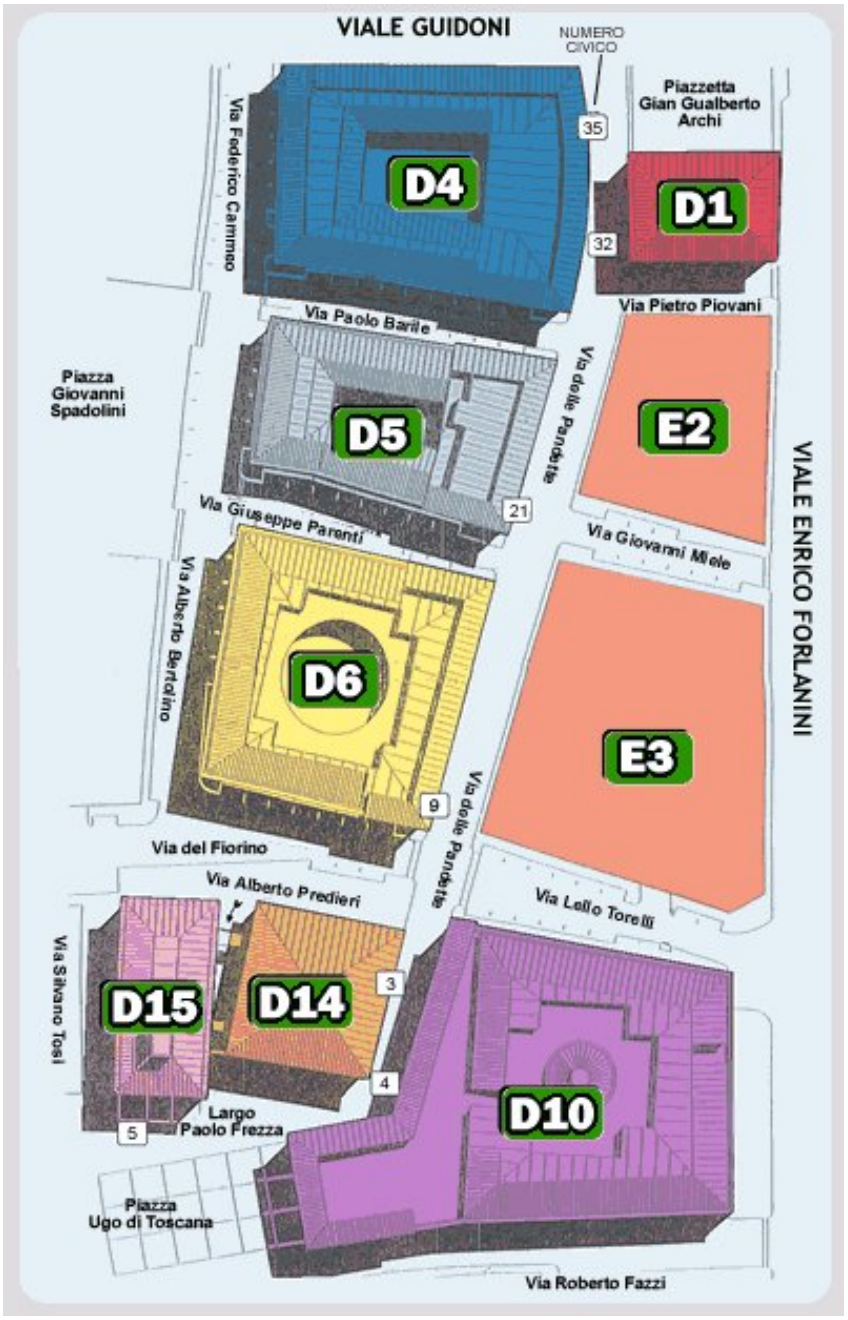


Fig. 2 Map of the “Polo delle Scienze Sociali”

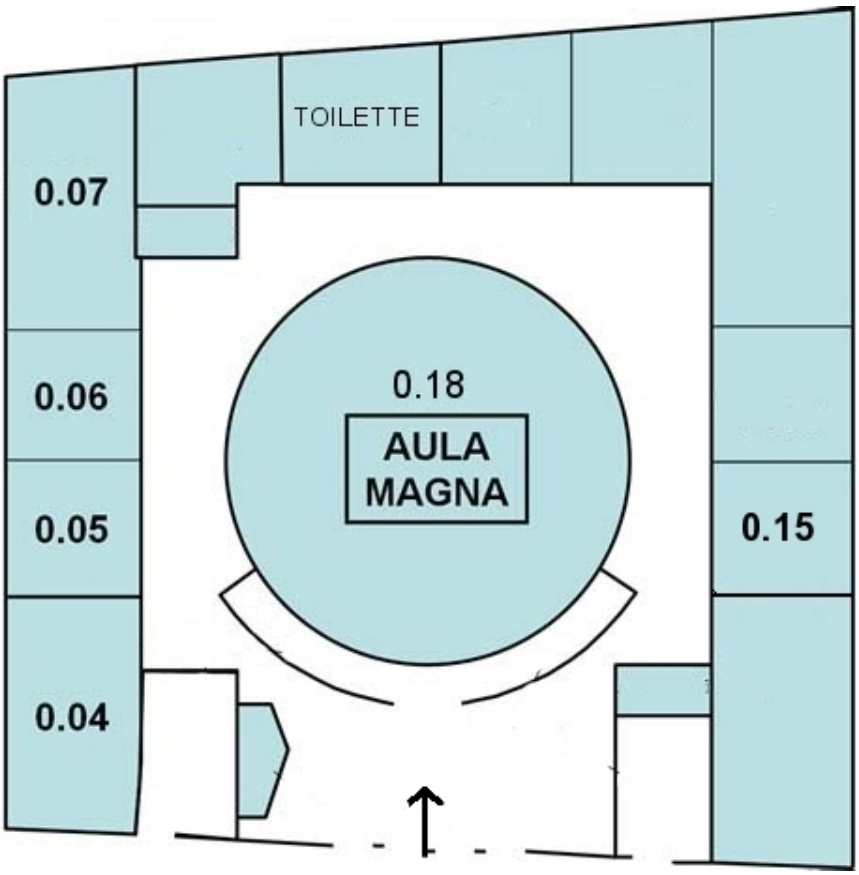


Fig. 3 Building D6, Ground floor - Rooms

All sessions will be held in building D6
At the “Polo delle Scienze Sociali”
Via delle Pandette, 9
FIRENZE

ROOMS
Room 0.04 seats 97
Room 0.05 seats 48
Room 0.06 seats 48
Room 0.07 seats 97
Room 0.15 seats 48
Room 0.18 seats 366

Abstracts

Contents

Plenary Sessions

Adaptive spectral clustering in molecular simulation 21
Marcus Weber, Susanna Röblitz, Fiete Haack

Critical issues and developments in microarray-based prediction 23
Anne-Laure Boulesteix

Knowledge extraction by investigating model uncertainty through predictive path modeling and probabilistic networks 25
Vincenzo Esposito Vinzi, Manaf Zargoush

Mixing mixtures of Gaussians 27
Roberto Rocci

Specialized Sessions

Specialized Session A
Challenges and Advances in Statistical Spatial Data Analysis

Urban knowledge discovery - Swiss population development of 15 decades 31
Martin Behnisch, Alfred Ultsch

Agronomic field grain property maps by geoadditive models: a comparison of different spatial correlation structures 33
Barbara Cafarelli, Alessio Pollice, Giovanna Jona Lasinio

Spatial model in small area estimation: an M-Quantile approach 35
Monica Pratesi, Stefano Marchetti, Caterina Giusti, Nicola Salvati

Specialized Session B
Robust Statistical Methods for Data Analysis: Methodological Innovations and Applications

Exploiting infinite variance through dummy variables in an AR model . . . 39
Giuseppe Cavaliere, Iliyan Georgiev

On robust estimation via pseudo-additive information 41
Davide Ferrari and Davide La Vecchia

Robustness versus consistency in ill-posed classification and regression problems 43
Robert Hable, Andreas Christmann

Specialized Session C
Web Data Mining

Monitoring the web sentiment, the Italian prime minister's case 47
Furio Camillo, Federico Neri

Web page importance ranking 49
Wolfgang Gaul

Evaluation metric for learning from imbalanced data based on asymmetric Beta distribution 51
Nguyen Thai-Nghe, Zeno Gantner, Lars Schmidt-Thieme

Specialized Session D
Opinion Mining and Preference Analysis

On the use of multivariate multiple regression models in the elicitation of consumer preferences 55
Giuseppe Giordano

A model-based approach for qualitative assessment in opinion mining . . . 57
Maria Iannario, Domenico Piccolo

Preference analysis for durable goods - Some remarks on data consistency and empirical results 59
Diana Schindler, Lars Luepke

Specialized Session E
Recent Developments in Recursive Partitioning Methods

An iterative fuzzy and time-dynamic clustering approach for time series . 63
Hans-Hermann Bock

Evaluating the performance of different sets of classifiers in multiclass learning 65
Claudio Conversano

Recursive partitioning of complex data for classification and regression trees 67
Roberta Siciliano

Specialized Session F
Statistical Signal Analysis

Musical instrument detection based on extended feature analysis 71
Markus Eichhoff, Igor Vatolkin, Claus Weihs

Auralization of auditory models 73
Klaus Friedrichs, Claus Weihs

Spatially constrained curve clustering: a hierarchical approach to signals analysis 75
Elvira Romano, Antonio Irpino

Specialized Session G
Classification in Systems Biology I

Penalized likelihood approaches for high-dimensional model selection . . . 79
Axel Benner

A shared components model to detect uncommon risk factors in disease mapping 81
Emanuela Dreassi

Making sense of large-scale proteomics datasets: myths, facts, and challenges associated with the analysis and statistical evaluation of protein ID and abundance data 83
Metodi V. Metodiev

Specialized Session H
Data Stream Mining

Summarizing and detecting structural drift from multiple data streams . . 87
Antonio Balzanella, Rosanna Verde

Information-based data stream summary 89
Fabrice Clérot, Pascal Gouzien

Tests for change detection based on incremental quantile estimation 91
Katharina Tschumitschew, Frank Klawonn

Specialized Session I
Data Visualization

Visualization of clustering comparisons with confusion matrices in Seurat 95
Alexander Gribov, Antony Unwin

Visualisation of cluster analysis results 97
Hans-Joachim Mucha, Hans-Georg Bartel

Some recent advances on dimension reduction for high-dimensional data . 99
Luca Scrucca

Specialized Session J
Classification in Systems Biology II

Phylogenomics as a standard technique for microbial taxonomy 103
Markus Göker

Boolean networks for modeling gene regulation 105
Hans A. Kestler

Graphical models for eliciting structural information 107
Federico M. Stefanini

Specialized Session K
Social Networks and Classification

Classification of personal networks using latent roles 111
Ulrik Brandes, Jürgen Lerner, Miranda J. Lubbers, Christopher McCarty, José Luis Molina, Uwe Nagel

Developments in blockmodeling 113
Anuška Ferligoj

Clustering social actors by using auxiliary information on relational data 115
Giuseppe Giordano, Maria Prosperina Vitale

Specialized Session L
Statistical Matching: Theory and Applications to Data Mining and Official Statistics

Matching two different topics: ecological inference and data fusion 119
Marcello D'Orazio, Marco Di Zio, Mauro Scanu

Displaying uncertainty in data fusion by imputation 121
Florian Meinfelder, Susanne Rässler

Data fusion for direct marketing 123
Peter van der Putten

Contributed Sessions

Contributed Session 1
Achievement and Research

Bibliometric indicators for statisticians: critical assessment in the Italian context 127
Francesca De Battisti, Silvia Salini

Measuring the effect of cultural capital on students' university achievement 129
Isabella Sulis and Mariano Porcu

Formative measurement model for academic reputation 131
Emma Zavarrone

Contributed Session 2
Applications in Economics

A regression clustering method for the prediction of the pro capita disposal income in municipalities 135
Paolo Chirico

Discrete and continuous time mover-stayer model for labour market in a small northern Italian area 137
Fabrizio Cipollini, Camilla Ferretti, Pero Ganugi, Mario Mezzanzanica

Electricity consumption and gross domestic product in the Italian regions 139
Antonio Angelo Romano and Giuseppe Scandurra

Contributed Session 3
Dimension Reduction

Variable selection in a predictive approach 143
Simone Borra, Agostino Di Ciaccio

Adaptive linear dimension reduction in a classification setting 145
Karsten Luebke, Claus Weihs

Dimensional reduction and clustering of functional data 147
Roberto Rocci, Stefano Antonio Gattone

Contributed Session 4
Environmental Data Analysis and Classification

Functional clustering of temperature and precipitation data for Italian climate zones determination 151
Edmondo Di Giuseppe, Giovanna Jona Lasinio, Massimiliano Pasqui, Stanislao Esposito

The application of M-function analysis to the geographical distribution of earthquake sequence 153
Eugenia Nissi, AnnaLina Sarra, Sergio Palermi, Gaetano De Luca

From a multivariate spatio-temporal array to a Multipollutant - Multisite Air Quality Index 155
Antonella Plaia, Mariantonietta Ruggieri, Francesca Di Salvo, Gianna Agró

Contributed Session 5
Applications in Demography

Analysis of the individual variability of sex ratio with hierarchical models 159
Silvia Bozza, Mario Di Bacco, Renzo Bigazzi, Sergio De Iasio

Clustering of population pyramids presented as histogram symbolic data 161
Simona Korenjak-Černe, Nataša Kejžar, Vladimir Batagelj

Modeling fertility and education in Italy: time-variant or invariant unobserved heterogeneity component? 163
Daniele Vignoli, Alessandra Mattei, Anna Gottard

Contributed Session 6
Handwriting and Web Analysis

The evaluation of handwriting evidence: multi-level models for determining authorship 167
Silvia Bozza, Franco Taroni, Raymond Marquis, Matthieu Schmittbuhl

Relative Linkage Disequilibrium in tracking web search patterns 169
Ron Kenett, Silvia Salini

Baysian classification tree for statistical user web-URL categories navigation pattern model 171
Klaokanlaya Silachan, Panjai Tantatsanawong, Chidchanok Lursinsap

Contributed Session 7
Applications to Sociology and Market

Integrating the spatial dimension into propensity score matching to evaluate regional impact of capital subsidies 175
Marusca De Castris, Guido Pellegrini

Asset ownership of the elderly across Europe: a multilevel latent class analysis to segment country and households 177
Omar Paccagnella and Roberta Varriale

Restructuring and innovations on the survey "capacity of collective tourist accommodation" and their impact on the process quality 179
Maria Teresa Santoro, Simona Staffieri

Contributed Session 8
Discrete Data

Generating ordinal data 183
Pier Alda Ferrari, Alessandro Barbiero

Gini measure: its decomposition proposal in the discrete case 185
Paolo Giudici, Emanuela Raffinetti

Adaptive discrete Beta kernel graduation of demographic data 187
Angelo Mazza, Antonio Punzo

Contributed Session 9
Classification Methods

Restricted Boltzmann machines for market basket analysis 191
Harald Hruschka

Classification of chunked data using Proximal Vector Machines and Singular Value Decomposition 193
Antonio Irpino, Mario Rosario Guarracino, Rosanna Verde

Robust kernel Fisher discriminant analysis with weighted kernels 195
Nelmarie Louw

Contributed Session 10
Time Series and Spatial Analysis

Forecasting and clustering beanplot time series 199
Carlo Drago and Germana Scepi

Local analysis of spatial relationships: a comparison of GWR and OLS method 201
Massimo Mucciardi, Pietro Bertuccelli

Classification of spatio-temporal series and hidden Markov models 203
Luigi Spezia

Contributed Session 11
Advances in Latent Class Modeling

A latent class version of the inverse probability-to-treatment weighted estimator for dynamic causal effects 207
Francesco Bartolucci, Fulvia Pennoni and Luca Pieroni

Finite mixture modeling of censored longitudinal data 209
Bettina Grün, Kurt Hornik

Determining the number of components in multilevel mixture (factor) models 211
Roberta Varriale, Jeroen K.Vermunt

Contributed Session 12
Clustering and Dissimilarities

Optimal length choice in top k ordered lists aggregation 215
Luisa Cutillo, Annamaria Carissimo, Diego di Bernardo

Recognition of ultrametrics and tree-metrics in optimal time 217
Bernard Fichet

A dissimilarity measure between two hierarchical clusterings 219
Isabella Morlini, Sergio Zani

Contributed Session 13
Customer Satisfaction and Conjoint Analysis

Unveiling non-linear relationships between perceived satisfaction and quality 223
Giuseppe Boari, Gabriele Cantaluppi

Ordinal logistic regression for the estimate of the response functions in the conjoint analysis 225
Amedeo De Luca

Individual Self Balancing Conjoint (ISBC): an adaptive design technique for choice based conjoint 227
Peter Kurz, Andrzej Sikorski

Contributed Session 14
Recent Trends in Classification

Clustering highly evolving data streams 231
Antonio Balzanella, Rosanna Verde, Yves Lechevallier

Extending SOM with efficient estimation of the number of clusters 233
Guénaël Cabanes and Younès Bennani

A classification approach for structure discovery in search spaces of combinatorial optimization problems 235
Daniel Cosmin Porumbel, Jin-Kao Hao, Pascale Kuntz

Contributed Session 15
Regression and Factor Analysis

Generalized extreme value regression in rare events 239
Raffaella Calabrese, Silvia Angela Osmetti

Ensemble procedures for more accurate Regression Trees with Moderating Effects 241
Gianfranco Giordano, Massimo Aria

On the use of bootstrap in factor analysis 243
John Ohrvik, Gabriella Schoier

Contributed Session 16
Biomedical Applications

Evolutionary Neural Networks to design synthetic proteins 247
Davide De March, Irene Poli

Understanding co-expression of co-located genes using a PCA approach . 249
Marion Ouedraogo, Frederic Lecerf, Sébastien Lê

Modeling delay to diagnosis for Amiotrophic Lateral Sclerosis: under reporting and incidence estimates 251
Irene Rocchetti

Contributed Session 17
Developments in Archetypal Analysis

Archetypal Analysis for interval valued data 255
Stefania Corsaro, Marina Marino

Archetypal analysis for prototype identification 257
Maria Rosaria D'Esposito, Francesco Palumbo, Giancarlo Ragozini

New features for archetypal analysis in R 259
Manuel Eugster, Friedrich Leisch

Contributed Session 18
Model Selection

Automatic detection of outliers in linear regression models: the Forward Search approach 263
Bruno Bertaccini, Franco Polverini

Estimating the Gap of the Forward Search via censored sampling 265
Danya Facchinetti, Silvia Angela Osmetti

Assessing model accuracy using a two-dimensional loss function 267
Roberto Savona, Marika Vezzoli

Contributed Session 19
Proximity Data and Hierarchies

Multi-objective genetic algorithm based clustering for dissimilarity data . 271
Laura Bocci

A new approach for analyzing a set of hierarchies 273
Marine Cadoret, Sébastien Lê, Jérôme Pagès

Extensions of modularity clustering 275
Andreas Geyer-Schulz, Michael Ovelgönne

Contributed Session 20
Professional Profiles

University graduate's job hunting: what helps to get the good one? 279
Marisa Civardi, Franca Crippa, Vincenzo Bagnardi

The detection of criticalities of graduates' jobs through Importance-Performance Analysis 281
Luigi Fabbris, Giovanna Boccuzzo

Statistical methods to describe professional profiles through competences and activities 283
Cristiana Martini

Contributed Session 21
Correspondence Analysis and Related Methods I

Joint Correspondence Analysis vs. Multiple Correspondence Analysis: a solution to an undetected problem 287
Sergio Camiz, Gastão Coelho Gomes

Measuring subcompositional incoherence in contingency tables and compositional data 289
Michael Greenacre

Correspondence Analysis in the case of outliers 291
Anna Langovaya, Hamdi Chouikha, Sonja Kuhnt

Contributed Session 22
Estimation Problems

Principal Stratification in sample selection problems with non normal error terms 295
Giovanni Mellace, Roberto Rocci

Some considerations on two-stage calibration estimators 297
Mario Montinaro, Ivan Sciascia

Automatic strategies of analysis for handling structurally missing occasions 299
Nadia Solaro

Contributed Session 23
Issues in Classification and Clustering

Model-based clustering of multistate data with latent change. An application with DHS data 303
José G. Dias

Issues on clustering and data gridding 305
Jukka Heikkonen, Domenico Perrotta, Marco Riani, Francesca Torti

Tree partitioning criteria across objects and predictors for data with a double stratification 307
Valerio A. Tutore, Valentina Cozza and Antonio D'Ambrosio

Contributed Session 24
Multivariate Analysis for Relational Data

Cluster analysis of multivariate relational data 311
Vladimir Batagelj

Multiple Correspondence Analysis for relational data 313
Domenico De Stefano and Giancarlo Ragozini

Analysis of multivariate event networks 315
Jürgen Lerner

Contributed Session 25
Rankings and Preferences

A Monte-Carlo study to evaluate value-added models for institutions' rankings 319
Bruno Arpino, Roberta Varriale

Ordinal models to assess media reputation 321
Paola Cerchiello, Paolo Giudici

Inference on the CUB model: an MCMC approach 323
Laura Deldossi, Roberta Paroli

Contributed Session 26
Correspondence Analysis and Related Methods II

Comparing mental maps: Obama vs. McCain 327
Simona Balbi, Michelangelo Misuraca, Emma Zavarrone

Assessing the response quality in ordered categorical data 329
Jörg Blasius

Adaptive factorial clustering for binary data 331
Alfonso Iodice D'Enza, Francesco Palumbo

Contributed Session 27
Methodology and Applications of Latent Class and Mixture Models

Statistic cognitive survey on Passito wine in Veneto region (Italy) from the consumer's point of view 335
Rosa Arboretti Giancristofaro, Stefano Bonnini, Elisa Grossule, Susanna Ragazzi, Luigi Salmaso

Covariate effects in multivariate latent growth models for the analysis of undergraduated student performances 337
Silvia Bianconcini, Silvia Cagnone and Paola Monari

On using item features to estimate parameters in IRT models 339
Mariagiulia Matteucci, Stefania Mignani, and Bernard P. Veldkamp

Contributed Session 28
Markov and Graphical Modeling

Model selection in latent Markov models: a simulation study 343
Michele Costa, Luca De Angelis

Bayesian analysis of longitudinal categorical data via latent Markov models 345
Silvia Pandolfi, Francesco Bartolucci and Alessio Farcomeni

On the identification of discrete graphical models with hidden nodes 347
Elena Stanghellini, Barbara Vantaggi

Contributed Session 29
Variable Structures

Hierarchical Factorial Classification of variables: methods and applications 351
Sergio Camiz, Jean-Jacques Denimal

Sensitivity analysis of composite indicators through Mixed Model Anova . 353
Cristina Davino, Rosaria Romano

A model for the clustering of variables taking into account external data . 355
Karin Sahmer

Contributed Session 30
Risk Analysis

The distribution of the stochastic dominance index for risk measurement 359
Silvia Facchinetti, Paolo Giudici, Silvia Angela Osmetti

Concentration measures for risk analysis 361
Silvia Figini, Paolo Giudici and Pierpaolo Uberti

Robust estimation and prediction for credit risk models 363
Silvia Figini, Luigi Grossi

Contributed Session 31
Miscellanea

Spatial clustering for local analysis 367
Federico Benassi, Chiara Bocci, Alessandra Petrucci

Symbolic tree for prognosis of localized osteosarcoma patient 369
Tae Rim Lee, Dae Geun Jeon, Edwin Diday

Index 371

Plenary Sessions

Adaptive spectral clustering in molecular simulation

Marcus Weber, Susanna Röblitz, Fiete Haack

Classical molecular simulation algorithms generate large high-dimensional data sets of molecular states. The aim of molecular simulation is to figure out the time-scales of molecular processes. Examples include binding processes of drug-like molecules and protein folding processes. In order to derive the time-scales from the simulation data, a clustering of the generated molecular states is needed. For example, in a simulation of a binding process some states have to be classified as "bonded drug molecule" and other states have to be classified as "non-bonded drug molecule" in order to derive the transition rate between these two classes Bujotzek and Weber (2009). However, from a physical point of view, transition rates do not exist in general for an arbitrary clustering of the simulation data Weber (2009); Weber and Kube (2008); Kube and Weber (2007). Only the spectral clustering algorithm PCCA+ (Robust Perron Cluster Analysis Deuflhard and Weber (2005)) provides a physically meaningful clustering of the data Weber (2009). Spectral clustering of high-dimensional large data sets is still an unsolved problem Haack (2009). The talk presents an adaptive spectral clustering method in order to circumvent the curse of dimensionality and in order to optimize the resolution of cluster boundaries locally. The optimization problem is based on an objective function which is not derived from graph theory. The optimization problem is formulated such that the time-scales of the simulation are preserved by the clustering Weber and Kube (2008). The result of the clustering is a decomposition of the molecular state space into fuzzy sets Weber (2009); Deuflhard and Weber (2005). In the talk it will be shown that fuzzy sets are not only used, because the result of PCCA+ is formulated in terms of membership functions. Membership functions are also helpful, because they are the

Marcus Weber,
Zuse Institute Berlin, Takustraße 7, D-14195 Berlin, Germany

Susanna Röblitz,
Zuse Institute Berlin, Takustraße 7, D-14195 Berlin, Germany

Fiete Haack
University Rostock, Joachim-Jungius-Str.10, D-18055 Rostock

key to a robust clustering algorithm from the viewpoint of perturbation theory. Furthermore, membership functions offer a different concept of hierarchical clustering, which perfectly suits to the situation of molecular simulation. To give an example: Sometimes simulation data can be interpreted as a 2-cluster situation (A) as well as a 3-cluster situation (B). Although, PCCA+ can be denoted as a hierarchical method, the 3rd cluster in (B) contains data from each of the 2 clusters in (A). In terms of membership functions, this is not paradox.

References

Haack F. (2009). *Repersentative Spectral Clustering for Large Data Sets applied to Gene Expression Data*, FU Berlin, Master's Thesis Bioinformatics.

Weber M. (2009). *A Subspace Approach to Molecular Markov State Models via an Infinitesimal Generator*, ZIB report 09-27.

Bujotzek A., Weber M. (2009). *Efficient Simulation of Ligand-Receptor Binding Processes Using the Conformation Dynamics Approach*, Journal of Bioinformatics and Computational Biology 7(5), 811-831.

Weber M., Kube S. (2008). *Preserving the Markov Property of Reduced Reversible Markov Chains*, Numerical Analysis and Applied Mathematics, Int. Conf. on Num. Analy. and Appl. Math. 2008, AIP Conference Proceedings, Kos 1048, 593-596.

Kube S., Weber M. (2007). *A Coarse-Graining Method for the Identification of Transition Rates between Molecular Conformations*, J. Chem. Phys. 126(2), 024103.

Deuflhard P., Weber M. (2005). *Robust Perron Cluster Analysis in Conformation Dynamics*, Lin. Alg. App. 398c, 161-184.

Critical issues and developments in microarray-based prediction

Anne-Laure Boulesteix

In the first part of this talk, I give an overview of state-of-the-art methods for classification with high-dimensional microarray data and error rate estimation in small sample settings. I also discuss current challenges and perspectives for statistical research including the design of very sparse prediction rules, the validation of the obtained rules and of their added predictive value, the incorporation of biological knowledge into the classification process, or the stability of prediction rules and gene lists as reviewed in Boulesteix and Slawski (2009).

In the second part of the talk, I present a novel approach for globally testing a large set of molecular covariates in prediction settings while adjusting for clinical covariates (Boulesteix and Hothorn, 2010). This permutation testing procedure that is based on boosting regularized regression turns out to perform particularly well in terms of power when there are few strong molecular covariates – a common situation in practice.

In the third part I illustrate the problem of optimization bias through an empirical study based on real-life microarray data sets (Boulesteix and Strobl, 2009). The strategy consisting to successively try many prediction methods in cross-validation and report only the smallest error rate in the paper induces a severe optimistic bias. In the study I present in this talk, a median error rate as low as 31% can be obtained based on non-informative predictors - just by ”fishing” for the lowest error rate after applying many classification methods (including penalized logistic regression, k-nearest-neighbors, random forests, etc). I also discuss potential solutions to this problem.

A related problem treated in the fourth part of the talk is the optimization bias in the context of methodological statistical research. When developing their new algorithms, researchers often adapt them sequentially to the data at hand in a trial-and-error approach. As a consequence, new algorithms often overfit the data set(s) used for their development. This obviously leads to an optimistic bias in the sense that

Anne-Laure Boulesteix,
Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Germany
e-mail: boulesteix@ibe.med.uni-muenchen.de

the superiority of the new method compared to existing approaches is substantially over-estimated. I illustrate this mechanism based on a concrete study on microarray-based classification incorporating priori knowledge from a biological database and outline the importance of validation (Jelizarow et al, 2010).

References

Boulesteix A.-L., Hothorn T. (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics*, 11, 78.

Boulesteix A.-L., Strobl C. (2009). Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology*, 9, 85.

Boulesteix A.-L., Slawski M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10, 556–568.

Jelizarow M., Guillemot V., Tenenhaus A., Strimmer K., Boulesteix A.-L. (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics* (conditionally accepted).

Knowledge extraction by investigating model uncertainty through predictive path modeling and probabilistic networks

Vincenzo Esposito Vinzi, Manaf Zargoush

1 Introduction

When studying complex systems the difficulty of analysis is mainly due to the complex network of hypothesized, but often hidden, and presumably causal (or at least predictive) relationships between tangible (i.e. manifest and directly observed) phenomena or intangible (i.e. theoretical and indirectly observed) concepts. It is somehow the problem of extracting knowledge from uncertain models rather than modeling uncertainty in a specific model defined on some a priori available information. The basic elements of causal networks (in a covariance-based framework) or predictive path models (in a component-based approach) are the manifest variables, the corresponding latent variables (or factors) and the network of dependence/causal relationships between the latter ones. Both the measurement model (manifest-latent links) and the structural model (latent-latent links) are usually specified according to theoretical hypotheses of the researcher and can be eventually (but only slightly) modified in case the statistical modeling of empirical data does not confirm the whole set of hypotheses thus providing a different or new evidence. Further knowledge may be extracted if induction by automatic learning is merged to the evaluation of probabilistic networks.

2 Main results

Causal or predictive modeling of relationships in a multi-block framework, based on classical covariance-based Structural Equation Modeling or its component-based

Vincenzo Esposito Vinzi,
ESSEC Business School of Paris, e-mail: vinzi@essec.edu

Manaf Zargoush,
ESSEC Business School of Paris, e-mail: zargoush@gmail.com

variant Partial Least Squares Path Modeling (Esposito Vinzi et al., 2010) and (Tenenhaus et al., 2005), may be limited for diagnosis by the theoretically hypothesized network of linear relationships in both the measurement and the structural submodels. In other words, when considered by these classical approaches, either unsuspected or nonlinear - even significant - relationships would be ignored. Bayesian probabilistic networks, which in turn are limited in differentiating between manifest and latent variables as well as between causal and undirected or even spurious relationships, instead are strong tools to eventually overcome the restrictions associated with the aforementioned classical approaches (Jensen and Nielsen, 2007) and (Neapolitan, 2003). As a consequence, these two approaches can be combined with the objective of discovering and validating a hidden network of relationships between manifest variables as well as between eventual underlying factors based on probabilistic causation. Automatic learning from manifest variables and related factors allows discovering unexpected relationships. This is followed by a probabilistic evaluation of different network candidates based on the adequacy between the data and the network as well as on the structural complexity of the network itself. Hidden factors may be then discovered again by induction (data analysis or mining tools). Finally, the outcome (or the hierarchy of outcomes) of this learning process becomes the input of a constrained statistical confirmatory analysis in the framework of structural equation or path modeling for validation and generalization. BayesiaLab (Bayesia, 2010) is undoubtedly the most representative software implementing Bayesian network capabilities in two major aspects: "(un)supervised automatic learning" and "analysis". While the unsupervised learning of general associations between variables ends up with global networks of relationships, the supervised learning is specifically intended to optimally characterize a target (manifest or latent) variable of interest to the researcher. An integrated approach can be applied either to manifest variables prior to estimating causality networks and predictive models or, a posteriori, to latent variable scores yielded by such models with different insights for both theory and practice in terms of diagnosis, prediction and further analysis of the discovered network.

References

Bayesia S.A. (2010). *BayesiaLab 4.6.8*, www.bayesia.com.

Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. EDS (2010). *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics, Heidelberg: Il Mulino.

Jensen, F.V., Nielsen, T.D. (2007). *Bayesian networks and decision graphs*, Springer Verlag.

Neapolitan, R.E. (2003). *Learning Bayesian Networks*, Upper Saddle River, NJ: Prentice Hall.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, 48 (1), 159-205.

Mixing mixtures of Gaussians

Roberto Rocci

1 Introduction: the problem

One of the most frequently used model in cluster analysis, is the finite mixture model. It is based on the hypothesis that the probability density of a multivariate observation is of the form

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i) \tag{1}$$

where $\mathbf{x}_i$ is a random vector sampled from a population formed by K subpopulations, with distributions $f_1, \ldots, f_K$, in proportions (prior probabilities) $p_1, \ldots, p_K$. Given a sample of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the model can be used to classify the observations into K classes, firstly, by computing the posterior probabilities $p_{k|i} = \Pr\{k|\mathbf{x}_i\}$, i.e. the probability that the i-th observation comes from component k, and then by using the MAP rule (Maximum A Posterior probability) to assign each observation to one of the components. Both model parameters and the number of components K are usually unknown and estimated from the data.

In model based cluster analysis, the mixture model is often applied by assuming a Gaussian distribution for the components and the observations assigned to a component are regarded as a cluster sampled from a specific subpopulation. Such interpretation is acceptable whenever the hypothesis that each group has a normal distribution is true. Actually, when one or more subpopulations have a distribution far from the Gaussian, the estimated number of components, say G, is usually greater than the true number K of subpopulations. In fact, each subpopulation can be well approximated by a finite mixture of Gaussians, hence two or more components, out of G, of the mixture are related to the same group. In this situation the problem of how to recover the clusters from the components arises.

Roberto Rocci,
Dipartimento SEFeMeQ, University of Rome "Tor Vergata", e-mail: roberto.rocci@uniroma2.it

2 Some solutions

In this work, we show how the aforementioned problem can be overcome. We assume an heterogeneous population formed by K subpopulations, or group, each having a distribution that can be described, or well approximated, by a finite mixture of Gaussians. In formulas, with obvious notation, in model (1) we set

$$f_k(\mathbf{x}_i) = \sum_{g=1}^{G} \frac{u_{kg}\pi_g}{\sum_h u_{kh}\pi_h} \phi_g(\mathbf{x}_i), \quad p_k = \sum_{g=1}^{G} u_{kg}\pi_g \tag{2}$$

where $\sum_g \pi_g \phi_g(\mathbf{x}_i)$ is a mixture of Gaussians and u_{kg} is equal to 1 if the g-th Gaussian component belongs to the k-th group and zero otherwise. In general, the consistent estimation of (2) is not possible if the labels of the K groups are unknown, because of lack of identification. However, we show how in particular applications the nature of the problem entails the specification of some constraints which permit the identification. This is done in the unity measure error (Di Zio et al., 2007) and Principal Stratification (Frangakis and Rubin, 2002) contexts. Whenever the application does not imply meaningful constraints, we first estimate a mixture of $G(>K)$ Gaussians, then, we combine the estimated components into K clusters by maximizing an appropriate criterion.
For fixed K, two different criteria are considered:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p_k \left(\frac{p_{k|i}}{p_k} - 1\right)^2 \qquad \chi^2 \text{ Distance}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p_{k|i} \log \frac{p_{k|i}}{p_k} \qquad \text{Kullback-Leibler divergence}$$

where it is straightforward to show that $p_{k|i} = \sum_g u_{kg}\pi_{g|i}$. The optimization problem is carried out by specific algorithms in a crisp, i.e. $u_{kg} \in \{0,1\}$, and fuzzy, i.e. $u_{kg} \in [0,1]$, framework. The two different criteria are compared with existing methods (see Hennig (2010) for a review and some new methods) on a theoretical and empirical basis, by using simulated and real datasets.

References

Di Zio, M., Guarnera, U., Rocci, R., (2007). A mixture of mixture models for a classification problem: the unity measure error. *Computational Statistics and Data Analysis*, 51, 5, 2573–2585.

Frangakis, C.E., Rubin, D.B., (2002). Principal stratification in causal inference. *Biometrics* 58, 191–199.

Hennig, C., (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4, 1, 3–34.

Specialized Session A

Challenges and Advances in Statistical Spatial Data Analysis

Urban knowledge discovery - Swiss population development of 15 decades

Martin Behnisch, Alfred Ultsch

1 Introduction

Spatial analysis is far from adequate handling the huge volumes of data and the growing complexity (Behnisch, 2009). In comparison to former approaches dealing with population (Bätzing and Dickhörner, 2001) the pool of data is examined in depth. The properties of distance measurements are considered in view of clustering.

2 Main results

The urban knowledge discovery approach provides the ability to identify and explain new patterns within a large amount of data. The initial idea was to use the development of population between 1850 and 2000 as a kind of overall indicator for the observable development of Swiss communities. Data relates to disposable official statistics (Swiss Federal Population Census). As an alternative to relative change calculation relative differences (Ultsch, 2008) are suggested to observe the population development of 15 decades. The investigation of distributions leads to intermediate results that three categories (“Loser”, “Typical”, and “Winner”) characterize Swiss communities in one decade. The modeled distribution and determined parameters (Mean, Standard Deviation and amount of communities by distribution) provide the identification of the degree of membership to a specific category. The Bayesian theorem offers advantages through its ability to formally incorporate prior knowledge into model specification via prior distributions and allows considering the variability. A pattern is defined as a unique profile of 15 categories (Winner, Typ-

Dr.-Ing. Martin Behnisch,
www.urban-data-mining.de/UDM.html, e-mail: Martin.Behnisch@urban-data-mining.de

Prof. Dr. Alfred Ultsch,
www.uni-marburg.de/fb12/datenbionik, e-mail: ultsch@informatik.uni-marburg.de

ical and Loser) over time. A classification process provides the discovery of multiple and partly unsuspected patterns over time. It was possible to identify 880 patterns in total. One pattern (15 $\times$ "Typical") is surprising because it is clearly representing the "Typical" Swiss population development over time. Another interesting aspect deals with patterns with one or zero "Non-Typical". They are characterizing more than 50 percent of all Swiss communities. All patterns and their (long-term) mean value of population are used for information optimization. The aim is to identify relevant patterns for the purpose of clustering. Such technique relates to the theoretical foundation of the Pareto 80/20-law (Ultsch, 2001). 65% of all communities belong to relevant patterns and about 85% of the Swiss population. By using the Euclidean Distance and Ward algorithm a clustering is realized of relevant patterns. It is based on growth indicators by three time intervals: 1850-1910, 1910-1950, and 1950-2000. The use of a pattern matrix (122 patterns $\times$ 15 decades) allows a visual comparison of contextual assumptions and confirms such division in a visual way. The result is a compact structure of eight specific population developments and their typical communities. The Emergent Self Organizing Map (Ultsch, 1999) is presented as an appropriate technique to visualize and verify the structure. A k-Nearest Neighbor classifier is constructed to allocate all Swiss communities to the existing partition. The process of class explanation provides the transition from data to knowledge. Since a partition of classes is realized it is important to foster the understanding of related spatial and aspatial characteristics. At first localization of classes is used for spatial verification and spatial reasoning. Secondly the detected class partition and other well-known typologies in Switzerland are compared using contingency tables in order to decide whether or not dependencies are significant. Third structure interpretation and validation in mind of the spatial analyst lead to knowledge about the Swiss communities. It triggers spatial abstractions and generates hypothesis that might be valuable for subsequent analysis.

References

Bätzing W., Dickhörner, Y. (2001). Die Typisierungen der Alpengemeinden nach 'Entwicklungsverlaufsklassen' für den Zeitraum 1870-1990. *Mitteilungen der Fränkischen Geographischen Gesellschaft*, 48, 273-303.

Behnisch M. (2009). *Urban Data Mining*, Karlsruhe: KIT Scientific Publishing.

Ultsch A. (1999). Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In: Kaski S., Oja E.(eds.) Kohonen Maps 1999, Maryland:Elsevier, 33-46.

Ultsch A. (2001). Proof of Pareto's 80/20 law and Precise Limits for ABC-Analysis. Technical Reports No. 30, Dept. of Mathematics and Computer Science, University of Marburg, Germany.

Ultsch A. (2008). Is log ratio a good value for measuring return in stock investments? In: Advances in Data Analysis, Data Handling and Business Intelligence 2008, Book of Short Papers, Meeting of the German Classification Society.

Agronomic field grain property maps by geoadditive models: a comparison of different spatial correlation structures

Barbara Cafarelli, Alessio Pollice, Giovanna Jona Lasinio

1 Introduction

Precision agriculture is an ecological management strategy based on the use of several sources of information to support decisions concerning agricultural applications with the aim of optimizing the use of soil and water resources and chemical inputs on a site specific basis. The adoption of soil management practices and natural resources conservation policies can take advantage of relevant spatial statistical methods, which can be helpful in developing a differential farm management calibrating different actions according to agricultural practices and soil conditions. In this paper we investigate an agricultural trial carried out on a 12-ha field cropped with durum wheat, located in Foggia, South-Eastern Italy during crop seasons 2005-2006 and 2007-2008. One-hundred georeferenced measurements of an indicator of durum wheat production (grain weight at harvest, GW), protein content of grain (P) and mean number of seeds for each ear of wheat (S) were taken for each crop season. After discarding samples with more than two missing values (resulting in 79 and 78 soil units for each crop season), exploratory data analysis showed that GW had a trend in the north-east direction and a marginal bell shaped distribution for both seasons. A geoadditive model was used to analyze the spatial distribution of grain weight and the nonlinear relations with other crop features. The proposed approach is a quick and effective method to predict the spatial distribution of grain weight by other grain features. The possibility of estimating nonlinear effects and variance components using standard mixed effects models softwares leads to the recommendation for a wider use of geoadditive models in precision agriculture.

Barbara Cafarelli,
Università degli Studi di Foggia, e-mail: b.cafarelli@unifg.it

Alessio Pollice,
Università degli Studi di Bari "Aldo Moro", e-mail: apollice@dss.uniba.it

Giovanna Jona Lasinio,
Università di Roma "Sapienza", e-mail: giovanna.jonalasinio@uniroma1.it

Table 1 Cross-validation results

		Full parametric	40 knots	30 knots	20 knots
Exponential	CV1	0.548	0.576	0.656	0.666
	CV2	1.175	1.185	1.228	1.232
	CV3	82.281	84.251	89.168	89.621
Spherical	CV1	0.548	0.547	0.605	0.588
	CV2	1.175	1.168	1.196	1.187
	CV3	82.275	82.690	87.514	86.126

2 Main results

Complex functional relations among *GW* measurements and other covariates led to the adoption of a semi-parametric approach (Kammann and Wand , 2003; Cafarelli and Pollice, 2008) based on the follwing model:

$$GW_{it} = \beta_0 + \beta_1 SW_{it} + f(P_{it}) + \beta_{\mathbf{x}} \mathbf{x}_{it} + S(\mathbf{x}_{it}) + \varepsilon_{it} \tag{1}$$

where x_{it} is the i-th spatial location during the t-th crop season with $i = 1, \ldots, 157$ and $t = 1, 2$. The term $f(\cdot)$ is a smooth function, $S(\cdot)$ is a zero-mean second order stationary Gaussian spatial random field. A penalized splines low-rank formulation (Kammann and Wand , 2003) of both $f(\cdot)$ and $S(\cdot)$ in (1) led to a mixed model representation of the geoadditive model. The exponential and spherical correlation structures and three different numbers of knots were considered for $S(\cdot)$. Also a fully parametric specification of the spatial correlation structure was used for comparison. The estimated models were compared by cross-validation statistics CV_1, CV_2 and CV_3 as suggested by Carroll and Cressie (1996). Results showed that the geoadditive model performed slightly better with the exponential spatial correlation than with the spherical one in terms of unbiasedness of the predicted values, accuracy of the mean squared prediction error and goodness of predictions. For the chosen model, the estimates of the linear component fixed effect and of the intercept resulted significant.

References

Cafarelli B., Pollice A. (2008). Geoadditive models for the analysis of the spatial distribution of soil salinity in a Sardinia coastal area. *Environmetrics*, 19, 742-750.

Carroll S. S., Cressie N. (1996). A comparison of geostatistical methodologies used to estimate snow water equivalent. *Water Resour. Bull.*, 32, 267-278.

Kammann E. E., Wand M. P. (2003). Geoadditive models. *Applied Statistics*, 52, 1-18.

Spatial model in small area estimation: an M-Quantile approach

Monica Pratesi, Stefano Marchetti, Caterina Giusti, Nicola Salvati

One possible approach to small area estimation when data are spatially correlated is to employ Simultaneous Autoregressive random effects models to define the Spatial Empirical Best Linear Unbiased Predictor. An alternative approach that incorporates the spatial information in the regression model is to use Geographically Weighted Regression (GWR). In GWR the relationship between the outcome variable and the covariates is characterised via local rather than global parameters.

In this paper we investigate GWR-based small area estimation under the M-quantile modelling approach. In particular, we specify an M-quantile GWR model that is a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This model is then used to define a bias-robust predictor of the small area characteristic of interest that also accounts for spatial association in the data. An important spin-off from applying the M-quantile GWR small area model is that it can potentially offer more efficient synthetic estimation for out of sample areas. We demonstrate the usefulness of this framework through both model-based as well as design-based simulation, with the latter based on a realistic survey data set.

The paper concludes with an illustrative application that focuses on small area estimation of the average equivalised income at LAU 1 and LAU 2 levels, Local Administrative Units 1 and 2, that is Provinces and Municipalities.

Monica Pratesi,
DSMAE, University of Pisa, e-mail: m.pratesi@ec.unipi.it

Stefano Marchetti,
DSMAE, University of Pisa, e-mail: s.marchetti@ds.unifi.it

Caterina Giusti,
DSMAE, University of Pisa, e-mail: caterina.giusti@ec.unipi.it

Nicola Salvati,
DSMAE, University of Pisa, e-mail: salvati@ec.unipi.it

References

Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2010). Small area estimation via M-Quantile Geographically Weighted Regression. *Paper under review*.

Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika* **93**, 255-268.

Fotheringham, A.S., Brundson, C. and Charlton, M. (2002). *Geographically Weighted Regression* West Sussex: John Wiley and Sons.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications* **17**, 113–141.

Specialized Session B

Robust Statistical Methods for Data Analysis: Methodological Innovations and Applications

Exploiting infinite variance through dummy variables in an AR model

Giuseppe Cavaliere, Iliyan Georgiev

A traditional role of impulse dummy variables in econometrics - to correct for a small number of observations that are not well described by a maintained model - has been complemented in recent research by situations where dummy variables are used in number proportional to the sample size and solely as a means to construct estimators or test statistics with desirable statistical properties. Such features are also present in this paper: we dummy out large time-series innovations drawn from a distribution with infinite variance although they perfectly fit the maintained infinite-variance model and although they are numerous. The justification for doing so are the properties of the resulting estimator.

Knight (1989, 1991) proved that in autoregressive (AR) time-series models with a unit root and innovations in an α-stable domain of attraction, $0 < \alpha < 2$, a class of M-estimators has better consistency rate than the OLS estimator and induces asymptotically Gaussian inference under the unit root hypothesis. Knight assumed that the M-estimators in question exist and are sufficiently close to the true parameter value but discussed no conditions or situations where these assumptions are actually satisfied (except of the trivial case where the M-estimator minimizes a strictly convex function).

In the proposed paper, on the other hand, we focus on a particular M-estimator which is not covered by Knight's setup but has the advantage of straightforward computability using dummy variables. It is attractive for practitioners, as it reflects the common practice of dealing with large residuals by including impulse dummies in the estimated auto-regression. We define the estimator by a particular iterative computational procedure rather than as a usual M-estimator in order to make sure that the object of theoretical study is in the same as the one obtained numerically. This settles the question of existence of our estimator. We provide conditions on the preliminary estimator used to initialize the iteration and on the threshold above

Giuseppe Cavaliere,
Università di Bologna

Iliyan Georgiev,
Universidade Nova de Lisboa

which observations are dummied out (allowing for the threshold to be estimated as a part of the iteration), such that:

- the dummy-based estimator is consistent at higher rates than the basic OLS estimator,
- an asymptotically Gaussian test statistic for the unit root hypothesis can be derived, and
- order of magnitude gains of local power obtain.

In our analysis of the iteration, the map defining the update of the estimates is studied as a random process. This involves the derivation of uniform asymptotic expansions of weighted empirical processes, similarly to Koul (2002). These allow us to approximate the sequence of iterates by an autoregressive process in "iteration time" and to deduce its properties of interest therefrom.

In previous studies on infinite-variance auto-regressions attention has often been restricted to the AR(1) model. We focus on the AR(1) model too, due to a property that seems specific to it, namely, a linear uniform asymptotic approximation to the map formalizing dummy-variable iteration. Although our main results, like consistency rates, asymptotic normality and convergence of dummy-variables iteration can be conjectured to carry over to the general AR(p) setting, a different approach to its study would be appropriate. In the AR(p) model, which can be transformed to have stationary regressors under the single unit root hypothesis, the coefficients of these regressors can be conjectured to be consistently estimable by dummy variable iteration at the rate established for other M-estimators (Davis et al., 1992) and for the computationally challenging ML estimator (Andrews et al., 2009). However, this rate equals the magnitude order of the stationary regressors, meaning that residuals are not uniformly infinitesimally close to true innovations. This is why the approximation would require modifications in the AR(p) case.

Finally, we document the relevance of our asymptotic finding by means of a Monte Carlo study.

References

Andrews B., Calder M. and Davis R. (2009). Maximum likelihood estimation for α-stable autoregressive processes, *Annals of Statistics*, forthcoming.

Davis R., Knight K. and Liu J. (1992). M-estimation for autoregressions with infinite variance, *Stochastic Processes and their Applications* 40: 145-180.

Koul H. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*. Berlin: Springer.

Knight K. (1989). Limit theory for autoregressive-parameter estimates in an infinite-variance random walk. *Canadian Journal of Statistics* 17:261-278.

Knight K. (1991). Limit theory for M-estimates in an integrated infinite variance process. *Econometric Theory* 7:200-212.

On robust estimation via pseudo-additive information

Davide Ferrari and Davide La Vecchia

Let $\mathscr{F}_\Theta = \{F_t, t \in \Theta \subseteq R^p\}$, $p \geq 1$ be a family of parametric distributions with densities $f_t$and let $\mathscr{G}$ be the class of all distributions G having density g, where G is the "true" distribution generating the data, regarded as close to some member of $\mathscr{F}_\Theta$. One way to estimate parameters is to minimize a data-based divergence measure between the candidate model F_t and an empirical version of G. By far, the most popular among minimum-divergence methods is maximum likelihood estimation, based on Kullback-Leibler divergence (Akaike, 1973). Although maximum likelihood is optimal when $G \in \mathscr{F}_\Theta$, deviations from the assumed model can severely affect its precision. On the other hand, traditional robust estimators tolerating such deviations, do not achieve first order efficiency for most parametric families. Beran (1977) puts forward a semi-parametric estimator based on minimization of Hellinger distance, which affords a large fraction of bad data, yet maintaining full efficiency. Basu and Lindsay (1994) extended Beran's approach by considering minimization of power divergences, a larger class of measures including Hellinger distance as a special case. The family of power divergences is defined by

$$D_q(f_t||g) = -\frac{1}{q}\int_{\mathscr{X}} L_q\left\{\frac{f_t(x)}{g(x)}\right\} g(x)dx, \tag{1}$$

where $L_q(u) = (u^{1-q}-1)/(1-q)$, and $q \in (-\infty,\infty)\setminus\{1\}$. Other notable divergences are special cases of (1): Kullback-Leibler divergence ($q \to 1$); twice Hellinger distance ($q = 1/2$); Neyman's Chi-square ($q = -1$) and Pearson's Chi-square ($q = 2$).

In principle, the above approaches seemingly settle the dispute between robustness and efficiency. In practice, however, minimization of (1) for a given data set requires kernel estimation of g. Therefore: (i) some nonparametric analysis for selecting the bandwidth is unavoidable, with nontrivial complications in multivari-

Davide Ferrari,
University of Modena and Reggio Emilia, Italy e-mail: davide.ferrari@unimore.it

Davide La Vecchia,
University of Lugano (USI), Switzerland e-mail: davide.la.vecchia@usi.ch

ate problems; (ii) the accuracy of the parameter estimator rests on the convergence rate of the kernel density smoother, which suffers from the course of dimensionality. In the present paper, we study a procedure for parameter estimation based on minimization of (1) which is fully parametric. This allows, at least in principle, to handle cases where $\dim(\mathscr{X})$ is moderate or large. Our approach has an information-theoretical flavor as it entails minimization for the generalized entropy function introduced by Havrda and Charvát (1967), sometimes called q-entropy:

$$H_q(f_t||g) = -\int_{\mathscr{X}} L_q\{f_t(x)\}\, g(x)dx. \tag{2}$$

Given the data $x_1,\ldots,x_n$, we find the parameter values by minimizing the stochastic counterpart of (2), $-\sum_{i=1}^{n} L_q\{f_t(x_i)\}$. We show that such a quantity is closely related to (1). The resulting parameter estimator is indexed by a single constant q tuning the trade-off between robustness and efficiency. If $q = 1$, we minimize the Kullback-Leibler divergence and the procedure is maximum likelihood estimation; if $q = 1/2$, the estimator minimizes a fully parametric version of the Hellinger distance. Intermediate choices give remarkable robustness and yet result in small efficiency losses, which tend to be negligible as the number of variables grows.

Asymptotic and infinitesimal robustness is addressed using influence and change-of-variance functions. Although both are useful tools to approximate worst-case bias and variance under contamination, to our knowledge, the change-of-variance function has been considered only in one-parameter location and/or scale problems for M-estimators (e.g., see Genton and Rouseeuw (1995)). Here, we devise a multi-parameter expression for the change-of-variance function and use it to study the mean squared error under contamination when $\dim(\Theta)$ is fairly large. The worst-case mean squared error is also employed for analytic selection of q (min-max approach).

References

Akaike, H. (1973). Information theory and an extension of the likelihood principle, in: 2nd international symposium of information theory.

Basu, A. and B. G. Lindsay (1994). Minimum disparity estimation for continuous models: Efciency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46, 683705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5, 445463.

Genton, M. and P. Rouseeuw (1995). The change-of-variance function of M-estimators of scale under general contaminations. *Journal of Computational and Applied Mathematics*, 64, 6980.

Havrda, J. and F. Charvát (1967). Quantication method of classication processes: a Concept of structural entropy. *Kibernetika*, 3, 3035.

Robustness versus consistency in ill-posed classification and regression problems

Robert Hable, Andreas Christmann

1 Introduction

There are a number of properties which should be fulfilled by a statistical procedure. First of all, it should be consistent, i.e., it should converge in probability to the true value for increasing sample sizes. Another crucial property is robustness, i.e., small model violations (particularly caused by small errors in the data) should not change the results too much. It is well-known from parametric statistics that there can be a goal conflict between efficiency and robustness. However, in many nonparametric statistical problems, there is even a goal conflict between consistency and robustness. That is, a statistical procedure which is (in a certain sense) robust cannot always converge to the true value. This is the case for so-called ill-posed problems. It is well-known in the machine learning theory that many nonparametric statistical problems are ill-posed. In particular, this is often true for nonparametric classification and regression problems. Here, we bring together notions and facts which are common in different fields, namely robust statistics and machine learning.

2 Main results

Many statistical estimation problems can be formalized in the following way: Let $\mathscr{P}$ be a set of probability measures on a metric space $\mathscr{Z}$ and let $\mathscr{F}$ be another metric space. It is assumed that one element $P_0 \in \mathscr{P}$ is the true probability measure and

Robert Hable,
University of Bayreuth, Department of Mathematics, D-95440 Bayreuth, Germany,
e-mail: Robert.Hable@uni-bayreuth.de

Andreas Christmann,
University of Bayreuth, Department of Mathematics, D-95440 Bayreuth, Germany,
e-mail: Andreas.Christmann@uni-bayreuth.de

the task is to estimate the value $T(P_0)$ of the functional $T : \mathscr{P} \to \mathscr{F}$. In parametric statistics, we typically have $\mathscr{P} = (P_\theta)_{\theta\in\Theta}$, $\mathscr{F} = \Theta \subset \mathbb{R}^k$ and $T(P_\theta) = \theta$ for all $\theta \in \Theta$. In nonparametric classification, $T(P)$ could be the whole function of conditional probabilities $x \mapsto P(Y = 1|X = x)$.

Ill-posedness has been connected with the statistical notion of qualitative robustness by Dey and Ruymgaart (1999): The problem of estimating $T : \mathscr{P} \to \mathscr{F}$ is *ill-posed* if T is not continuous with respect to weak convergence of probability measures. For the mathematical definition of qualitative robustness, see Hampel (1971) and Cuevas (1988, Def. 1). Roughly speaking, qualitative robustness implies that an estimator is hardly affected by small errors in many data points and large errors in only a small fraction of the data set. A sequence of estimators $T_n : \mathscr{Z}^n \to \mathscr{F}$, $n \in \mathbb{N}$, is *(universally) consistent* for the problem of estimating $T : \mathscr{P} \to \mathscr{F}$ if it converges in probability to the true value $T(P_0)$ for every true distribution P_0. Theorem 1 follows from Hampel (1971, Lemma 3) and Cuevas (1988, Theorem 1):

Theorem 1. *If the problem of estimating T is ill-posed, then no sequence of estimators T_n, $n \in \mathbb{N}$, can be simultaneously consistent and qualitatively robust.*

This goal conflict has recently been investigated for support vector machines (SVMs) in Hable and Christmann (2009). Depending on the choice of the regularization parameter, SVMs can either be qualitatively robust or consistent. Though the fact that many problems are ill-posed is well-known in machine learning theory, the implications concerning robustness have hardly received any attention: a procedure which is universally consistent cannot be qualitatively robust. This is somewhat contrary to a result from Poggio et al. (2004) which says that, for the method of empirical risk minimization, universal consistency is equivalent to their notion of stability even though empirical risk minimization typically is ill-posed. This shows that, in case of an ill-posed empirical risk minimization problem, no statistical procedure can be both qualitatively robust and stable in their sense. This indicates that the notion of stability (common in machine learning theory) and the notion of qualitative robustness are quite conflicting even though stability is sometimes considered as some kind of a robustness property in machine learning theory.

References

Cuevas A. (1988). Qualitative robustness in abstract inference. *Journal of Statistical Planning and Inference*, 18:277–289.

Dey A.K., Ruymgaart F.H. (1999). Direct density estimation as an ill-posed inverse estimation problem. *Statistica Neerlandica*, 53(3):309–326.

Hable R., Christmann A. (2009). Qualitative robustness of support vector machines. arXiv:0912.0874v1. *Submitted*.

Hampel F.R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896.

Poggio T., Rifkin R., Mukherjee S. and Niyogi P. (2004). General conditions for predictivity in learning theory. *Nature*, 428:419–422.

Specialized Session C

Web Data Mining

Monitoring the web sentiment, the Italian prime minister's case

Furio Camillo, Federico Neri

The Web is a huge virtual space where to express individual opinions and influence any aspect of life, with implications for companies and political parties alike. Nowadays the Web is becoming more and more crucial in the competitive political arena: politicians and their consultants can monitor electors suggestions or claims, or the perception they might have about leaders statements and decisions, by analyzing blogs, newsgroups and newspapers in real time. Spin doctors can analyze and capitalize on the explosion of individual opinions expressed online, in order to design populist measures and increase dramatically their leaders consensus.

The revolution in information technology is making open sources more accessible, ubiquitous, and valuable, making Open Source Intelligence and Web Sentiment Analysis at less cost than ever before. The world today is really in the midst of an information explosion. Anyway, the availability of a huge amount of data in Internet and in all the open sources information channels has lead to the well-identified modern paradox: an overload of information has meant, most of the time, a no usable knowledge. In fact, all the electronic texts are - and will be - written in various native languages, but these documents are relevant even to non-native speakers. The most valuable information is often hidden and encoded in pages which for their nature are neither structured, nor classified. Nowadays everyone experiences a mounting frustration in the attempt of finding the information of interest, wading through thousands of pieces of data. The process of accessing all these raw data, heterogeneous both for source, type, protocol and language used, transforming them into information, is therefore inextricably linked to the concepts of automatic textual analysis and synthesis, hinging greatly on the ability to master the problems of multilinguality.

Despite much progress in Natural Language Processing (NLP), the field is still a long way from a full Natural Language Understanding (NLU). In fact, understand-

Furio Camillo,
University of Bologna

Federico Neri,
University of Pisa

ing requires processing and knowledge that goes beyond parsing and lexical lookup and that is not explicitly conveyed by linguistic elements. Contextual understanding is needed to deal with the omissions. Ambiguities are a common aspect of human communication. Speakers are cooperative in filling gaps and correcting errors, but automatic systems generally are not. A mixed qualitative and quantitative approach can bridge this gap. This paper describes a Knowledge Mining study performed on over 1000 blog posts or news articles about the Italian Prime Minister Silvio Berlusconi, by using the iSyn Semantic Center platform.

iSyn Semantic Center is a content enabling system that provides deep semantic information access and dynamic classification features for large quantities of distributed multimedia data.

Web page importance ranking

Wolfgang Gaul

We start with data that could be / have been used for Web page importance calculations, discuss examples known from the literature and focus on centrality measures for linkage data dependent problems. Since the publication of PageRank many algorithmic approaches for the computation of Web page importance rankings have appeared (see, e.g., HITS (Hyperlink Induced Topic Search), SALSA (Stochastic Approach for Link Structure Analysis), OPIC (Online Page Importance Computation) for early techniques). A majority of algorithms is based on PageRank and / or suggests fast computations via linear system approaches or eigenvector solutions. Personalisation, query dependence, topic sensitivity, dynamic changes of the Web, Web decomposition aspects, and online / offline computation possibilities belong to the characteristics that Web page importance valuations should take into account. Thus, Web data analysis of the kind considered in this paper is a salient ingredient for Web intelligence enhancement strategies. Against this background a new algorithm is proposed that uses both degree and query relevance information w.r.t. Web pages.

Wolfgang Gaul,
Institut für Entscheidungstheorie und Unternehmensforschung, Karlsruher Institut für Technologie (KIT), Campus Süd, Postfach 6980, 76049 Karlsruhe, e-mail: wolfgang.gaul@kit.edu

Evaluation metric for learning from imbalanced data based on asymmetric Beta distribution

Nguyen Thai-Nghe, Zeno Gantner, Lars Schmidt-Thieme

1 Introduction

Class imbalance is one of the reasons degrading the classifier's performance. To evaluate the classifiers in this case, the area under the ROC curve (AUC) is commonly used. Recently, Hand (2009) has shown that using AUC is equivalent to averaging the misclassification loss over a cost ratio distribution, which depends on the score distributions. Since the score distributions depend on the classifier, this means that, the AUC evaluates different classifiers using different metrics. To overcome this incoherence, "H measure", which uses a symmetric Beta distribution to replace the implicit cost weight distribution in the AUC, was proposed.

When learning from imbalanced data, misclassifying a minority example is much more serious than misclassifying a majority example. To take the different misclassification costs into account, we propose using metric based on asymmetric Beta distribution such as $Beta(x;4,2)$ (B42) derived from H. Moreover, He and Garcia (2009) indicate two problems for future researches in class imbalance: Need of **standardized evaluation** and need of **uniform benchmark** as well as **large datasets** (Jamain and Hand (2009)). The contributions of this work are *i)* to propose metric for evaluating the classifiers on imbalanced data, *ii)* to introduce large and imbalanced benchmark datasets for systematic studies, and *iii)* to investigate the influence of class imbalance on the behavior of classifiers when learning from large datasets. Initial results show that B42 gives results different from AUC when comparing logistic regression (ℓ_2-LR) with ℓ_2-SVM (base). We also find out why undersampling (RUS) and weighting (W) methods do not work well for large imbalanced data.

Nguyen Thai-Nghe,
University of Hildesheim, Germany. e-mail: nguyen@ismll.uni-hildesheim.de

Zeno Gantner,
University of Hildesheim, Germany. e-mail: gantner@ismll.uni-hildesheim.de

Lars Schmidt-Thieme,
University of Hildesheim, Germany. e-mail: schmidt-thieme@ismll.uni-hildesheim.de

2 Main results

Obviously, B42 places more weights on the minority examples so it has some statistically significant (level=0.05) results different with AUC and H. Table 1 summarizes results evaluated on 36 datasets having minority class examples from 49.9% to 0.02%. B42 evaluates ℓ_2-LR better than ℓ_2-SVM 10 significant times (bold number) but AUC disagrees those results while the reverse is 6 times.

Table 1 B42 disagrees with AUC 16 significant times (left) and with H 8 times (right) out of 36

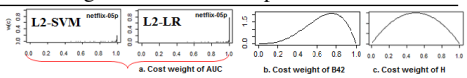
		B42	
		Sign. diff.	No-sign. diff.
AUC	Significant diff.	7	***6***
	No-sign. diff.	***10***	13

		B42	
		Sign. diff.	No-sign. diff.
H	Significant diff.	11	***2***
	No-sign. diff.	***6***	17

Moreover, AUC evaluates ℓ_2-LR better than ℓ_2-SVM on 3 groups (Table 2) while B42 shows that when the imbalanced ratio increases, ℓ_2-LR shifts from win (3/9/0) to tie (3/6/3) results. Thus, asymmetric B42 should be taken into account when evaluating on imbalanced data. The last column displays an example of cost weight distributions implicitly used in AUC (*different cost distributions for ℓ_2-LR (at 1.0) and ℓ_2-SVM on the same "netflix-05p"*) and explicitly used in B42 and H.

Table 2 Paired t-test (win/tie/lose) on 3 groups (12 datasets/group): ℓ_2-LR vs. ℓ_2-SVM (base)

	B42	AUC	H	Cost weight distribution samples
Group 3: %minor. 30-50	3/9/0	5/7/0	3/9/0	
Group 2: %minor. 5-30	6/5/1	3/9/0	4/8/0	
Group 1: %minor. 0.02-5	3/6/3	5/7/0	4/7/1	



To analyze the influence of class imbalance on the behavior of classifier (ℓ_2-LR), we select 11 large datasets and compare when learning on original data (as a baseline) with RUS, W, and feature selection (FS). We find out that RUS and W do not work well on large data while FS gives positive results. The (win/tie/lose) t-test of B42 are (7/4/0), (0/2/9), and (5/2/4) for FS, RUS, and W respectively.

References

Hand D.J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve, *Machine Learning*, 77(1), 103–123.

He H., Garcia E.A. (2009). Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

Jamain A., Hand D.J. (2009). Where are the large and difficult datasets?, *Advances in Data Analysis and Classification*, 3(1), 25–38.

Specialized Session D

Opinion Mining and Preference Analysis

On the use of multivariate multiple regression models in the elicitation of consumer preferences

Giuseppe Giordano

1 Introduction

The aim of this paper is to define an integrate scheme of analysis where different kinds of information are retrieved in suitable data matrices according to: *i*) the role of the variables (active/illustrative; response/predictor), *ii*) the kind of data (quantitative/qualitative) and *iii*) the purpose of the analysis (exploratory/confirmatory). We start from the formulation of the traditional Multivariate Linear Regression model that considers a set of multiple responses linearly related to a set of predictors. The model assumes that the expected values of the responses are linearly related to the predictors through an estimated coefficients matrix. According to the quantitative or categorical nature of data, different specification of the model can be derived (e.g. MANOVA, MANCOVA, Seemingly Unrelated Regression, and so on (Timm, 2002)).

The interest in this formulation arises from the specification of the metric approach to Conjoint Analysis, (Green and Srinivasan, 1990), which is one of the most famous techniques to analyse preference models at individual level. The data structure of the Conjoint Model is fully consistent with a multiple regression analysis with multiple preference responses revealed by a set of judges.

In this paper it is proposed a unified approach to investigate the typical data structure of Conjoint Analysis models in the framework of Exploratory Data Analysis. Firstly, the Factorial Approach to Conjoint Analysis is recalled, (Lauro et Al., 1998) and then, we introduce an *L*-shaped data structure where several variables are available as external information (socio-economic variables observed on the judges).

By taking into account the kind and the role of information retrieved in each data matrix (either metric or dummy variables), we aim at showing how to mine preference data from this peculiar data structure. We derive a unique formulation to

Giuseppe Giordano,
Dept. of Economics and Statistics, University of Salerno, Via Ponte Don Melillo, 84084 - Fisciano (Salerno), Italy. e-mail: ggiordan@unisa.it

investigate both the relationships among the different sets of predictors. (Giordano and Scepi, 1999). Namely, we express the preference response variables as a function of two sets of predictors: inner and outer arrays in the language of Design of Experiments (DoE).

The first dataset we consider is the matrix of predictor variables, $\mathbf{X}(n, k)$, where each of the k columns retrieve the factor-levels assigned to n runs (stimuli). In the DoE $\mathbf{X}$ is the Design Matrix, and it holds information coded according to suitable expanded indicators (dummy) variables. We look at the matrix $\mathbf{X}$ as an internal informative structure or *Inner Array*. The second data source is constituted by a different set of predictor variables. Let us define $\mathbf{Z}$ (q, g) the matrix where the vector rows identify q predictors, while the g columns are the statistical units. Thus, $\mathbf{Z}$ will be considered an external informative structure or *Outer Array*.

The matrix $\mathbf{Y}$ (n, g) holds two-ways response variables. It is related by rows to the matrix $\mathbf{X}$ and by columns to the matrix $\mathbf{Z}$. The three data matrices defined above can be combined into two multivariate multiple regression models that fit separately (eq. 1) the linear relationships between response variables and factors in $\mathbf{X}$, and between the response variables in $\mathbf{Y}'$ and the variables in $\mathbf{Z}'$:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E_1}\,(a); \quad \mathbf{Y}' = \mathbf{Z}'\mathbf{D} + \mathbf{E_2}\,(b) \tag{1}$$

In model (1.b) statistical units and variables in $\mathbf{Y}$ have exchanged their role. In the analysis of the preference models, this allows to take into account the effects of the stimulus' attributes and the Judges' characteristics on the elicited preference. Finally, an inter-relationships coefficient matrix will be derived according to the (2.a) and (2.b):

$$\mathbf{B}' = \mathbf{Z}'\theta + \mathbf{E_3}\,(a); \quad \mathbf{D}' = \mathbf{X}\theta' + \mathbf{E_4}\,(b) \tag{2}$$

which share the common analytic solution (3) whose properties will be discussed.

$$\theta = (\mathbf{ZZ}')^{-}\mathbf{ZY}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-} \tag{3}$$

References

Giordano G., Scepi G. (1999), Different informative structures for quality design, Journal of the Italian Statistical Society, 8, 2-3, pp.139-149.

Green P. E., Srinivasan V. (1990), Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice, The Journal of Marketing, Vol. 54, 4, pp. 3-19.

Lauro C.N., Giordano G., Verde R. (1998), A Multidimensional Approach to Conjoint Analysis, Applied Stochastic Model and Data Analysis, 14, Wiley, pp.265-274.

Timm N. H. (2002) Applied Multivariate Analysis, Springer, New York.

A model-based approach for qualitative assessment in opinion mining

Maria Iannario, Domenico Piccolo

1 Introduction

Data mining is an increasing area of interest where the collection of information on a large amount of data is an ordinary task and the synthesis of substantive concepts is a relevant topic. Since information is quite different with respect to origin and content, data mining is a research area where a high degree of specialization is required. Thus, we limit ourselves to consider opinions that are the result of explicit questions and are often related to liking, disliking or indifference positions with regard to a specific object. Specifically, we focus on ordinal data involving evaluation, comparison or perception: from a statistical point of view, we are modelling ratings and/or rankings of "objects" performed by a large number of subjects in a consistent way thanks to a Likert scale.

In this area, it is often difficult to summarize and visualize hundredths or thousands of expressed preferences on several objects. Moreover, explorative measures (based on descriptive indexes) miss to enhance the relevant components of preferences and evaluations, that is the latent constructs which cause the final expression. Current approaches relate such expressions to subjects' covariates by means of log-odds of distribution function. Alternatively, we will focus on mixture models of discrete probability distributions as proposed by Piccolo (2003); D'Elia and Piccolo (2005).

The main idea is that opinions are formed by reasoning and/or perception and both produce some evaluation, that is an ordinal assessment (of qualitative nature) about the object. When we ask people to express a preference on a rating scale we

Maria Iannario,
Department of Statistical Sciences, University of Naples Federico II,
e-mail: maria.iannario@unina.it

Domenico Piccolo,
Department of Statistical Sciences, University of Naples Federico II,
e-mail: domenico.piccolo@unina.it

are compelling them to force such opinion (quite often a fuzzy one) into prefixed ordinal bins. This process manifests itself as a combination of a personal *feeling* accompanied by an intrinsic *uncertainty*. Feeling is a sentiment (generated by several causes and well fitted by a unimodal distribution) whereas uncertainty is mostly related to circumstances of the choice. Any response will result the joint effect of such random variables with a varying degree; then, it seems consistent to model the given response as a weighted mixture of two discrete distributions (CUB models). Formally, for given m categories, we interpret rated opinions $(r_1, r_2, \ldots, r_n)'$ as realizations of a discrete random variable R with probability distribution:

$$Pr(R = r) = \pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi) \left(\frac{1}{m}\right), \qquad (1)$$

where $r = 1, 2, \ldots, m$. It is well defined on the parametric space $\Omega(\pi, \xi) = \{(\pi, \xi) : 0 < \pi \leq 1; 0 \leq \xi \leq 1\}$ and it is identifiable Iannario (2010) for $m > 3$. Moreover, it is possible to relate the parameters $(\pi, \xi)'$ to uncertainty and feeling components, respectively. After a brief discussion of the main features of a CUB model (interpretation, inferential issues and fitting measures), we show how to visualize such estimated models by means of a one-to-one correspondence between parameter values and $\Omega(\pi, \xi)$. Thus, we may assess and summarize expressed preferences as a collection of points and test the possible effect of covariates, space and time circumstances, cohorts' variability, and so on. We also provide a more accurate and comprehensive description of relationships in clustered data. A program in R is available for performing such tasks Iannario and Piccolo (2009). Further developments and extensions of such models are useful to fit more elaborate conditions, for instance when a *shelter effect* modifies the distribution of responses or if mixed effects of covariates are present. Some evidence on real data sets will confirm the effectiveness of the proposed approach.

Acknowledgements. Authors thanks CFEPSR, Portici, for the availability of structures. The research has been partly supported by a grant from MIUR (code 2008WKHJPK-PRIN2008).

References

D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.

Iannario M. (2010). On the identifiability of a mixture model for ordinal data, *METRON*, LXVIII, 87–94.

Iannario M., Piccolo D. (2009). A program in R for CUB models inference, Version 2.0, available at *http://www.dipstat.unina.it/CUBmodels1/*

Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.

Preference analysis for durable goods - Some remarks on data consistency and empirical results

Diana Schindler, Lars Luepke

1 Introduction

Today the internet has established as an important vehicle for online trading and chatting. Customers intensively use new forms of communication, such as forums, blogs, discussion groups and consumer review platforms like Epinions.com and Cnet.com to express their opinions and thoughts about nearly everything of everyday life, and in particular about products and services. The respective customer feedback data has emerged as an efficient base for observing, analyzing and tracking products and services. Unstructured full text reviews, semi-structured pro/con summaries and structured object rating data collected from appropriate online platforms indicate what consumers think about a product as a whole and about its different attributes. A still widely undervalued advantage of this type of marketing data is the fact that online reviews are written free of any external force (Liu, 2008; Na and Thet, 2009) and provide a promising base for customer preference elicitation. Though, there is a rapidly growing literature on opinion mining techniques, comparatively little research has been devoted to innovation-oriented preference analysis and the related question of data consistency. However, the lower the consistency of a set of reviews is, the less reliable the elicited preferences are.

Diana Schindler,
Department of Business Administration and Economics, Bielefeld University,
e-mail: dschindler@wiwi.uni-bielefeld.de

Lars Luepke,
Department of Business Administration and Economics, Bielefeld University,
e-mail: lluepke@wiwi.uni-bielefeld.de

2 Proceeding and Results

This paper, among others, focuses on the question of how to verify the consistency of online product reviews in the research setting sketched above. We start from the assumption that almost every consumer reviews platform nowadays asks its contributors to provide multiple information (full text, pro/con summary and rating) about the product(s) of interest. However, due to different external factors like low involvement, willingness to joke and faking, the consistency of online reviews is not guaranteed per se (Liu, 2010). Furthermore, sentiment classification is widely used to identify the sentiment value (positive, negative or neutral) of a text (Esuli and Sebastiani, 2006), but little is known about the ”correlations” between sentiment value, rating and pro/con summary.

Against this background, we suggest a new approach which combines negative binomial regression (Decker and Gnibba-Yukawa, 2009) with correlation analysis and NLP based sentiment classification in order to measure the overall degree of consistency and to identify possible sources of inconsistency. Furthermore, it will be shown how consistent opinion data can be used to elicit useful information for innovation-oriented preference analysis. The applicability and benefits of this approach are demonstrated using a large data set on digital cameras containing more than 15.000 online reviews, all of them being equivalent in structure and composition.

References

Decker R., Gnibba-Yukawa K. (2009). Konsumentenforschung im Web 2.0 - Analyse von Online-Rezensionen zur kundenorientierten Produktgestaltung. Marketing ZFP, 31 (2), 117-136.

Esuli A., Sebastiani F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, Genova.

Liu B. (2010). Sentiment Analysis and Subjectivity. In: Indurkhya N., Damerau F. J.. Handbook of Natural Language Processing. Chapman and Hall, London, 627-666, (forthcoming).

Liu B. (2008). Web data mining - Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin.

Na J.C., Thet T.T. (2009). Effectiveness of Web Search Results for Genre and Sentiment Classification. Journal of Information Science, 35 (6), 709-726.

Specialized Session E

Recent Developments in Recursive Partitioning Methods

An iterative fuzzy and time-dynamic clustering approach for time series

Hans-Hermann Bock

1 Introduction

We consider the problem of clustering n observed time series $\mathbf{x}_k = \{\, x_k(t) \mid 0 \le t \le T \,\}$, $k = 1,...,n$, with observation interval $\mathscr{T} = [0,T]$ and $x_k(t) \in R^d$, into a suitable number m of clusters each one comprizing time series with a 'similar' structure. There are many ways of describing the similarity of time series, based either on suitable dissimilarity measures followed by classical clustering techniques, or on probabilistic models for class-specific processes followed by a maximum likelihood or mixture approach.

Here we present a clustering approach that builds a dynamic fuzzy classification for the n given time series. It comprizes (1) a set of m class-specific parametric 'prototype functions' $\mu(t;\vartheta^{(i)})$ in R^d, $t \in \mathscr{T}$, with given form (e.g., constant, regression function, ARMA model), but unknown parameters $\vartheta^{(i)} \in \Theta$ $(i = 1,...,m)$ and $\theta := (\vartheta^{(1)},....,\vartheta^{(m)})$; (2) for each time point t a fuzzy m-clustering $\mathscr{U}(t)$ of the n time series, formalized by the membership matrix $\mathscr{U}(t) := (u_{ik}(t))_{m\times n}$ where $u_{ik}(t)$ is the degree of membership of series k to cluster i at time t, with constraints $u_{ik}(t) \ge 0$ and $\sum_{i=1}^{m} u_{ik}(t) = 1$ for $i = 1,...,m$, $k = 1,...,n$, $t \in [0,T]$; (3) thereby cluster memberships $u_{ik}(t)$ may change in time, but with the constraint that these functions do not vary 'too fast' or chaotically.

A suitable clustering criterion is proposed and minimized w.r.t. the matrix-valued clustering function $\mathscr{U} := \{\mathscr{U}(t) | t \in \mathscr{T}\}$ and the prototype parameters in θ. This is possible after approximating the membership functions by a finite linear combination of window functions in analogy to a finite-element approach by Horenko (2009). The corresponding optimization algorithm proceeds by a k-means like iteration method that alternates between determining an optimum parameter configuration for the prototypes, and solving a quadratic optimization problem under linear constraints for the unknown coefficients of the memberships.

Hans-Hermann Bock,
Institute of Statistics, RWTH Aachen University, e-mail: bock@stochastik.rwth-aachen.de

2 Main results

Clustering consists here in minimizing the integrated clustering criterion (mean deviation from the prototypes):

$$G(\mathscr{U},\theta) := \int_0^T \sum_{i=1}^m \sum_{k=1}^n u_{ik}(t) \cdot D(x_k(t), \mu(t;\vartheta_i))dt + \varepsilon^2 \cdot \int_0^T \sum_{i=1}^m \sum_{k=1}^n |\dot{u}_{ik}(t)|^2 dt \quad (1)$$

with respect to $\mathscr{U}$ and θ where $D(x,\mu)$ is a dissimilarity measure between values x,μ (from R^d) (e.g., $D(x,\mu) := ||x-\mu||, ||x-\mu||^2, ...$), the latter sum a penalty term that forces a suitably smooth behaviour of the memberships $u_{ik}(\cdot)$, and $\varepsilon > 0$ a given weight. The difficult infinite-dimensional minimization w.r.t. the continuous functions $u_{ik}(\cdot)$ is reduced to a standard finite-dimensional, and even quadratic, minimization problem by (a) segmenting $[0,T]$ into M time segments (intervals) $S_1 = [0,\tau_1), S_2 = [\tau_1,\tau_2), S_3 = [\tau_2,\tau_3), ..., S_M = [\tau_{M-1},T]$; (b) using for the membership functions $u_{ik}(\cdot)$ a linear 'finite-element' decomposition $u_{ik}(t) = \sum_{j=1}^M \phi_{ik}^{(j)} \cdot v_j(t)$ with suitable segment-specific window functions $v_1(t), ..., v_M(t)$, (unknown) coefficients $\phi_{ik}^{(j)}$, and coefficient vectors $\phi_{ik} := (\phi_{ik}^{(1)}, ..., \phi_{ik}^{(M)})' \in R^M$; (c) thereby considering the 'fuzziness constraints' $\phi_{ik}^j \geq 0, \sum_{j=1}^M \phi_{ik}^j = 1 \quad$ for all i,j,k. The clustering criterion (1) then reduces to the semi-quadratic form

$$G(\mathscr{U},\theta) := \sum_{i=1}^m \sum_{k=1}^n a_k'(\vartheta_i)\ \phi_{ik} + \varepsilon^2 \cdot \sum_{i=1}^m \sum_{k=1}^n \phi_{ik}' H \phi_{ik} \quad (2)$$

with a fixed, known matrix $H \in R^{MM}$ (from segment-specific integrals of the weight functions and their derivatives), and parameter-dependent vectors $a_k(\vartheta_i) := (a_k^{(1)}(\vartheta_i), ..., a_k^{(M)}(\vartheta_i)'$ with $a_k^{(j)}(\vartheta_i) := \int_0^T v_j(t) \cdot D(x_k(t), \mu(t;\vartheta_i))\ dt$. Alternating minimization of (2) w.r.t. the parameter vector $Theta$ (typically a classical parameter estimation problem) and the coefficient vectors ϕ_{ik} (quadratic minimization under linear and positivity constraints) yields a k-means like algorithm that can be performed with standard software packages.

References

Douzal-Chouakria A., Naghabushan P.N. (2007). Adaptive dissimilarity index for masuring time series proximity, *Advances in Data Analysis and Classification*, 1, 5–22.

Horenko I. (2009a). Finite element approach to clustering of multidimensional time series, *SIAM J. on Scientific Computation*.

Horenko, I. (2009b). On clustering of non-stationary meteorological time series, *Dynamics of Atmospheres and Oceans*, DOI:10.1016/j.dynatmoce.2009.04.003.

Mörchen F. (2006). Times series knowledge mining. Ph.D. thesis, University of Marburg, Germany. Grich & Weiershuser, Marburg, ISBN 3-89703-670-3.

Evaluating the performance of different sets of classifiers in multiclass learning

Claudio Conversano

Multiclass learning requires a classifier to discriminate instances (objects) among several classes of an outcome (response) variable. Since general purpose learning algorithms do not handle this multiclass classification problems in a proper manner, most of the studies do not address the whole problem; rather, a relatively small set of classifiers is trained to choose among these. To obtain a *K-class classifier*, the most common approach is to construct a set of binary classifiers each trained to separate one class from the rest (the so-called *One Versus the Rest*) and to combine them by using a model averaging criteria as, for example, voting (Hastie and Tibshirani, 1998; Furnkranz, 2002; Cutzu, 2003). A somewhat similar approach is *Error-Correcting Output Coding* (Dietterich and Bakiri, 1995). Here, a large number of binary classification problems is generated by splitting the original set of classes into two subsets and by training a binary classifier for each possible dichotomization. The final classification derives from a synthesis of the results obtained from each binary classification example which are stored in a decoding matrix composed of $\{\pm 1\}$. Alternatively, other approaches directly cast the multiclass classification problem into an objective function that simultaneously estimates a multiclass classifier as, for instance, in the Weston and Herbich's (2000) approach based on Support Vector Machines (SVM). This approach lacks of feasibility in some situations because the optimization of the objective function is sometimes difficult to obtain.

All these approaches are valuable because they allow the research community to develop better learning methods and evaluate them in a wide range of applications, but it is worth realizing that an important stage is missing: in some situations, analyzing such complex datasets requires the results to be easily interpretable. Taking the last consideration into account, this paper presents an approach for multiclass learning that tries to balance two different goals: The first one is to extract information on how predictors are associated with the response variable (i.e., knowledge extraction), whereas the second goal is to develop a model able to predict as well as possible the response value of a future observation (i.e., prediction rule).

Claudio Conversano,
Department of Economics, University of Cagliari, e-mail: conversa@unica.it

I refer to the *Sequential Automatic Search of Subset of Classifiers* (SASSC) algorithm (Conversano and Mola, 2010; Mola and Conversano, 2008) as an approach able to find the right compromise between knowledge extraction and good prediction. SASSC is an iterative algorithm that works by building a taxonomy of classes in an ascendant manner: this is done by the solution of a multiclass problem obtained by decomposing it to several r-nary problems ($r \geq 2$) in an agglomerative way. The algorithm begins with every class representing a singleton object. At each of the $K-1$ steps the closest two (least dissimilar) classes (or subsets of classes) are merged into a single *superclass*, producing one less class at the next higher level. As a measure of dissimilarity, SASSC uses the generalization error obtained with cross-validation once that two classes (or superclasses) are aggregated. The best aggregation is the one producing the lowest generalization error. Consequently, an overall measure of quality for the superclasses obtained in each iteration is provided by the weighted average of the generalization errors obtained from the classifiers trained on each superclass. The final result is a hierarchical aggregation of subsets of observations (response classes). Each of those subsets can be referred to a proper partition of the initial set of observations. The user can choose a final partition according to an overall measure of accuracy. Previous implementation of SASSC uses classification trees as base classifiers. In this paper, I extend the range of possible classifiers to be used in the iterations of SASSC and compare either the predictive performance or the interpretation issues related to the use of each set of classifiers.

References

Conversano C., Mola F. (2010) Detecting subset of classifiers for multi-attribute response prediction. In Lauro C.N., Greenacre M.J., Palumbo F. (eds): *Studies in classification, data analysis, and knowledge organization*, Springer, Berlin-Heidelberg, 225–232.

Cutzu F. (2003) Polychotomous classification with pairwise classifiers: a new voting principle. In Windeatt T., Roli F. (eds): *Multiple classifier system, Proceedings of the 4th international workshop MCS 2003*, Springer-Verlag, New York, 115–124.

Dietterich T.G., Bakiri G. (1995) Solving multi-class learning problems via error-correcting output codes, *J. Artif. Intell. Res.*, 2, 263–286.

Furnkranz J. (2002) Round Robin classification, *J. Mach. Learn. Res.*, 2, 721–747.

Hastie T.J., Tibshirani R.J. (1998) Classification by pairwise coupling, *Ann. Stat.*, 26(1), 451–478.

Mola F., Conversano C. (2008) Sequential automatic search of a subset of classifiers in multiclass learning. In Brito P. (ed): *Compstat 2008 Proceedings in Computational Statistics*, Springer, Berlin, 291–302.

Weston J., Herbich R. (2000) Adaptive margin support vector machines. In Smola A.J., Bartlett P.L., Scholkopf B., Schuurmans D. (eds): *Advances in large margin classifiers*, MIT press, Cambridge (MA), 281–295.

Recursive partitioning of complex data for classification and regression trees

Roberta Siciliano

Recursive partitioning methods are the core of supervised classification and non parametric regression methods, also known as classification and regression trees, tree-based methods, binary segmentation, decision trees (Hastie *et al.* , 2001). Input data consists of a response variable of numerical or categorical type and a set of predictors (numerical, ordinal, categorical) observed on a large sample of objects. No probability assumptions are required, whereas a data-driven procedure based on computational intensive algorithms are necessary for both learning and predicting tasks.

In learning process, a recursive partitioning of the sample provides an exploratory tree (either classification or regression tree), that can be analyzed to understand the dependence relationship between the response variable (of categorical or numerical type) and the predictors (of any type). In binary trees, at any internal node, a predictor generates the splitting variable (i.e., dummy variable) to discriminate the objects falling into the left subnode from those falling into the right subnode. Usually, the best splitting variable is chosen among all possible splits generated by the predictors such to maximize the decrease of impurity of the response variable within the two subnodes, where the impurity is a measure of deviance or variation for numerical responses (i.e., regression tree) and a measure of heterogeneity or entropy for categorical responses (i.e., classification tree). Terminal nodes include disjoint and homogeneous subgroups of objects, defining a partition of the starting group of objects with respect to the response variable.

In predicting process, the aim is to predict a response class/value of new objects in which only the measurements of predictors are known. One approach is to define a set of nested pruned subtrees by removing at turn the most unreliable branches and then select the most accurate subtree for new cases. As measure of accuracy the error rate for classification problem and the mean square error for regression problem are evaluated on independent test sample as well as according to a cross-validated

Roberta Siciliano,
Department of Mathematics and Statistics, University of Naples Federico II, Monte Sant'Angelo, Via Cinthia, I-80126 Naples, Italy. e-mail: roberta@unina.it

sampling procedure. Another approach consists in ensemble methods, such as boosting and bagging algorithms. This results in decision tree-based rules for new cases. The main difference between the two approaches is that in the former there is no possibility to obtain a decision tree structure such to interpret the classification or regression assignment of new objects while getting a better accuracy.

A one-step methodology to provide an exploratory tree that can be used also for prediction in efficient way has been recently developed (Siciliano *et al.* , 2008). Main idea is to obtain a prediction tree structure for known and unknown cases based on the optimal partitioning of known objects and a posterior prediction model for unknown objects. New concepts of optimal partitioning and posterior prediction modeling have been defined.

This result belongs to the library of methods provided in the last twenty years by the research unit in Naples, where all methods have been implemented in MATLAB environment providing the specialized software "Tree Harvest". Computational enhancements in recursive partitioning are also important to accelerate the prediction process based on ensembles. Main issues of recursive partitioning methods in Tree Harvest are to deal with complex data where standard methods fail as well as to exploit some additional information through modeling data in the partitioning criteria definition. Some examples of complex data are preference rankings as response, multilevel or hierarchically structured data, three-way data where instrumental variables play a role in the analysis, web data where a kind of survival analysis of the time spent on a web page should be considered, data with missing values. Recent results in recursive partitioning methods for classification and regression trees dealing with complex data will be overviewed.

References

Conversano S., Siciliano R. (2009). Incremental Tree-Based Imputation with lexicographic ordering, *Journal of Classification*, 26(3), 361–379.

Hastie T., Friedman J.H., Tibshirani R. (2001). *The Elements of Statistical Learning*. Springer Verlag.

Pecoraro M., Siciliano R. (2008). Statistical Methods for User Profiling in Web Usage Mining, in: Handbook of Research on Text and Web Mining Technologies, edited by Min Song and Yi-Fang Brook Wu, chapter XXII, IDEA Group. Inc., Hershey:USA.

Siciliano R., Aria M. and D'ambrosio A. (2008). Posterior Prediction Modelling of Optimal Trees, in: Proceedings in Computational Statistics 18th Symposium Held in Porto, Portugal, Brito, Paula (Ed.), Springer-Verlag, 323–334.

Tutore V.A., Aria M. and Siciliano R. (2007). Conditional classification trees using instrumental variables. *LNCS 4723, Advances in Intelligent Data Analysis*, Springer-Verlag, 163–173.

Specialized Session F

Statistical Signal Analysis

Musical instrument detection based on extended feature analysis

Markus Eichhoff, Igor Vatolkin, Claus Weihs

1 Introduction

In recent years musical instrument recognition as part of automatic music classification consists of considering low-level features such as e.g. zero-crossing, MFCC, spectral bands, spectral centroids, etc. as you can find in Theimer et al. (2008).

High-level features are more abstract than low-level features. They might be derived from or constructed with low-level features, but in every case they use additional information like musicological or physical knowledge of the musical instrument (Fletcher et al. , 1998) or the piece of music. An example of high-level features you find in Abesser (2008).

Because timbre is characterized by the distribution of overtones, it's interesting to analyze the so called pitchless periodograms (similar to chromagrams) as a high-level feature (Weihs and Ligges, 2003). This and other features like the amplitude envelope or spectral envelope (also in combination with each other) are used to classify single tones played by piano, guitar, string and reed instruments of the McGill University Master Samples database and other databases and therefore to optimize the instrument recognition. Classification with different statistical methods is carried out and its results are compared.

Aside of this approach the software AMUSE - a JAVA framework that unifies different MIR tasks - is used for low-level feature extraction and classification by amongst others evolutionary algorithms (Vatolkin et al., 2009). Both approaches are compared.

Markus Eichhoff, Claus Weihs,
Chair of Computational Statistics, TU Dortmund, Germany,
e-mail: eichhoff@statistik.tu-dortmund.de, weihs@statistik.tu-dortmund.de

Igor Vatolkin,
Chair of Algorithm Engineering, TU Dortmund, Germany,
e-mail: igor.vatolkin@cs.tu-dortmund.de

2 Main results

Some mean-misclassification errors (mmce) of the piano/guitar classification are shown in Table 1 below.

A 10-fold cross-validation is used to train on 275 piano and 270 acoustic guitar and e-guitar tones. The test-set consists of other 4039 guitar and 1070 piano tones. Statistical classification methods such as LDA, MDA, regression trees (Rpart), decision trees (RandomForest) or boosting methods (AdaBoost) are applied.

Some methods like Rpart, AdaBoost or RandomForest use hyperparameters. The mean-missclassification error is then the lowest mean-misclassification error of all combinations of hyperparameters.

Table 1 Classification[a] of piano and guitar tones (error rates (mmce) in %)

Methods	Training	Testing
MDA	6.79	12.98
Rpart	14.89	10.80
AdaBoost	0.73	2.00
RandomForest	4.23	3.95

[a] Used feature combination in this case: LPC-simplified spectral envelope + non-windowed MFCC's + Pitchless Periodogram

References

Abesser J., Dittmar C., Grossmann H. (2008). Automatic genre and artist classification by analyzing improvised solo parts from musical recordings. In: *Proceedings of Audio Mostly 2008*, Pitea, Sweden.

Fletcher N.H., Rossing T.D. (1998). *The Physics of Musical Instruments*, New York: Springer.

Theimer W., Vatolkin I., Eronen A. (2008). Definitions of Audio Features for Music Content Description, Algorithm Engineering Report TR08-2-001, Technical University Dortmund, Germany.

Vatolkin I., Theimer W., Rudolph G. (2009). Design and Comparison of Different Evolution Strategies for Feature Selection and Consolidation in Music Classification. In: *Proceedings of the 2009 IEEE Congress on Evolutionary Computation (CEC)*, Trondheim, Norway.

Weihs C., Ligges U. (2003). Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: Kopiez R., Lehmann A.C., Wolther I., Wolf C., editors, In: *Proceedings of the 5th Triennial ESCOM Conference*, Hannover University of Music and Drama, Germany, 332-335.

Auralization of auditory models

Klaus Friedrichs, Claus Weihs

1 Introduction

Computational auditory models describe the transformation from acoustic signals into spike firing rates of the auditory nerves by emulating the signal transductions of the human auditory periphery.

The inverse approach is called auralization, which can be useful for many tasks, such as quality measuring of signal transformations or reconstructing the hearing of impaired listeners. There have been a few successful attempts of auditory inversion but they all deal with relatively simple auditory models (Slaney et al., 1994), (Feldbauer et al., 2005).

In recent years more comprehensive auditory models have been developed, like the one of Meddis and Sumner (Sumner et al., 2002), which simulates nonlinear effects in the human auditory periphery. For this model an analytical inversion is not possible. Instead we propose an auralization approach using statistical methods.

2 Main results

To simplify the problem in our first studies only harmonic tones are examined. First statistical features of the spike firing rates are calculated. By using these features all possible frequency components of the original tone can be discovered. In the following step decision trees are used to classify which of these frequencies are really

Klaus Friedrichs,
Chair of Computational Statistics, TU Dortmund, Germany,
e-mail: friedrichs@statistik.tu-dortmund.de

Claus Weihs,
Chair of Computational Statistics, TU Dortmund, Germany
e-mail: weihs@statistik.tu-dortmund.de

partial tones. Afterwards just the power of each partial tone has to be estimated to resynthesize the original signal.

We used a training set, consisting of 30,000 randomly generated harmonic tones, to create the decision trees. A 10-fold cross validation was used to calculate their error rates. The results of the error rates are frequency dependent and located in the range of 0.1% and 1.3%. For a final test we built a test set of 1,000 harmonic tones, each consisting of up to 8 partial tones. A number of 967 of these signals was detected correctly, for the rest only a small error occurred.

In summary it means that using statistical methods seems to be a promising approach for auralization of auditory models. Future studies have to analyze the power estimation of the partial tones.

References

Feldbauer C., Kubin G. and Kleijn W.B.(2005).Anthropomorphic Coding of Speech and Audio: A Model Inversion Approach. In: EURASIP Journal on Applied Signal Processing, Volume 2005.

Hohmann V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. In: Acta acustica / Acustica, 88(3), p. 433-442.

Jepsen M.L., Dau T. and Ewert S. (2006). A model of the normal and impaired auditory system. Academic dissertation, Technical University of Denmark.

Slaney M., Naar, D. and Lyon R.F. (1994). Auditory model inversion for sound separation. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Adelaide, Australia.

Sumner C.J., O'Mard L.P., Lopez-Poveda E.A. and Meddis R., A revised model of the inner-hair cell and auditory nerve complex. In: Journal of the Acoustical Society of America.

Spatially constrained curve clustering: a hierarchical approach to signals analysis

Elvira Romano, Antonio Irpino

1 Introduction

In this paper we propose a strategy for clustering geographically referenced curves considered as functional data (Ramsay (2005)) and usually called spatial functional data (Giraldo and Dedicado and Comas and Mateu (2009)). This is a typical case of data recorded continuously over a period, coming from sensors, which are located in space.

In the statistical literature of the last years, many clustering methods for functional data have been proposed (Tarpey (2005)). To our knowledge few methods have been developed for taking into account spatial dependence among functional data. Two main approaches can be distinguished: model based (Serban (2009))and exploratory approaches (Giraldo and Dedicado and Comas and Mateu (2009)).

Spatial clustering can also be viewed as a regionalization procedure. Indeed, the algorithm searches for groups of data that are homogeneous not only for the observed value but also for the localization (the spatial constrain).

Here we propose a strategy for regionalization of spatial functional data based on a hierarchical clustering approach. The strategy consists in the following steps:

- constructing the Voronoi tessellation of the sensors in space;
- computing a contiguity matrix associated with the spatial tessellation structure;
- performing a hierarchical clustering on a dissimilarity matrix constrained by a contiguity matrix. Since a contiguity matrix can also be considered as a representation of undirected graph, we consider, as proposed by Guo (2008), two constrained strategies: a First Order Constrained Linkage Clustering, and a Full Order Constrained Linkage Clustering. In the first approach it is assumed that two sensors can be grouped if they are contiguous, in the second approach, two

Elvira Romano, Antonio Irpino,
Dipartimento di Studi Europei e Mediterranei, Facoltà di Studi Politici, Seconda Università degli Studi di Napoli, Via del Setificio 15, 81100 Caserta e-mail: elvira.romano@unina2.it, e-mail: antonio.irpino@unina2.it

sensors are aggregated if they have weight the minimum shortest path compared to other pair of sensors.

In order to decide which clusters should be combined a measure of dissimilarity between curves is required. We achieve this by using an euclidean metric. Along with evaluating the strategy by several linkage criteria, respectively minimum, avarage and maximum linkage coherently with Guo (2008), we also utilize a ward linkage, since we use an Euclidean distance among functions.

A classic Quality Partition index, that is the ratio among the Between variability and total Variability, is finally used for evaluating the different strategies.

2 Conclusion and main results

The proposed strategy differently from the existing approaches is able to work without any restrictive assumption of isotropy and stationarity, that is, the mean and variance functions are not necessarily constant and the covariance does not depend only on the distance between sampling points of the Spatial functional process. According to our procedure the statistical properties of each region can be explored after the clusters have been discovered.

References

Giraldo, R., Delicado, P., Comas, C., Mateu, J.(2009). Hierarchical clustering of spatially correlated functional data. Technical Report, http://www.ciencias.unal.edu.co/estadistica/reporte02.pdf, Universidad Nacional de Colombia Bogotá, COLOMBIA.

Guo, D. (2008).Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP),*Int. Journal of Geographical Information Science*,22, 7, 801–823, Bristol, PA, USA.

Jiang, H., Serban, N. (2009), Large Scale Clustering of Dependent Curves. http://www2.isye.gatech.edu/ nserban/clusteringresearch.html

Ramsay, J. and B. W. Silverman (2005). *Functional Data Analysis*(Second ed.). New York: Springer.

Tarpey, T. and Kinateder, K. (2003), Clustering functional data, *Journal of Classification*, 20, 93-114.

Specialized Session G

Classification in Systems Biology I

Penalized likelihood approaches for high-dimensional model selection

Axel Benner

One important topic of current research on observational and especially prognostic factor studies is the development of methods that can be employed to analyse high-dimensional data, where the number of explanatory variables is much larger than the number of observations. This is mainly driven by the requirements of biomedical applications such as DNA microarrays. The major problem of analyzing such data is the danger of overfitting. Methodological challenges arise in using large sets of covariates, e.g. patients gene expression profiles, to predict survival endpoints on account of the large number of variables and their complex interdependence.

The aim of this talk is to show how penalized regression models can be employed to analyse high-dimensional data. This include linear, logistic and proportional hazards regression models.

We illustrate the different approaches using real data examples from clinical microarray studies including gene expression data. The results will be discussed with respect to the prediction error and interpretability of the results.

One important topic of current research on observational and especially prognostic factor studies is the development of methods that can be employed to analyse high-dimensional data, where the number of explanatory variables is much larger than the number of observations. This is mainly driven by the requirements of biomedical applications such as DNA microarrays. The major problem of analyzing such data is the danger of overfitting. Methodological challenges arise in using large sets of covariates, e.g. patients gene expression profiles, to predict survival endpoints on account of the large number of variables and their complex interdependence.

The aim of this talk is to show how penalized regression models can be employed to analyse high-dimensional data. This include linear, logistic and proportional hazards regression models.

We illustrate the different approaches using real data examples from clinical microarray studies including gene expression data. The results will be discussed with respect to the prediction error and interpretability of the results.

Axel Benner, German Cancer Research Center, Heidelberg, e-mail: benner@dkfz.de

A shared components model to detect uncommon risk factors in disease mapping

Emanuela Dreassi

A statistical model for jointly analysing the spatial variation of incidences of three (or more) diseases, with common and uncommon risk factors, is introduced. Deaths for different diseases are described by a logit model for multinomial responses (multinomial logit or polytomous logit model). For each area and confounding strata population the probabilities of death for each cause are estimated. A specific disease, the one having a common risk factor only, acts as the baseline category. The log odds are decomposed additively into shared (common to diseases different by the reference disease) and specific structured spatial variability terms, unstructured unshared spatial terms and confounders terms to adjust the crude observed data for their effects. Disease specific spatially structured effects are estimated; these are considered as latent variables denoting disease-specific risk factors.

The model is presented with reference to a specific application. We considered the mortality data relative to oral cavity, larynx and lung cancers in the municipalities of Region of Tuscany (Italy). All these pathologies share smoking as a common risk factor; furthermore, two of them (oral cavity and larynx cancer) share alcohol consumption as a risk factor. Lung cancer acts as the baseline category. All studies suggest that smoking and alcohol consumption are the major known risk factors for oral cavity and larynx cancers; nevertheless, in this paper, we investigate the possibility of other different risk factors for these diseases, or even the presence of an interaction effect between known risk factors but with differential spatial patterns for oral and larynx cancer.

We assume that $y_{ij} = (y_{ij1}, \ldots, y_{ijk}, \ldots, y_{ijK})\prime$ deaths for the i-th municipality, j-th age-class and k-th disease, follows a multinomial distribution with parameters m_{ij} and probability vector $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijk}, \ldots, \pi_{ijK})\prime$, where $m_{ij} = \sum_{k=1}^{K} y_{ijk}$ and $\sum_{k=1}^{K} \pi_{ijk} = 1$. We consider a polytomous logit model considering an extension to more than two categories of the proportional mortality model (Dabney *et al.*, 2005); each category probability is modeled as

Emanuela Dreassi,
Dept. of Statistics, University of Florence, Italy, e-mail: dreassi@ds.unifi.it

$$\pi_{ijk} = \phi_{ijk} / \sum_{r=1}^{K} \phi_{ijr} \text{ where each log-odd } \log(\phi_{ijk}) = \alpha_k^\star + a_{jk}^\star + u_{ik}^\star + v_{ik}^\star \quad (1)$$

is decomposed additively into a disease-specific intercept $\alpha_k^\star$ (representing overall difference between k-th disease and K-th reference disease), $a_{jk}^\star$ a time-structured term by age and disease representing difference between k-th disease and reference category, and structured $u_{ik}^\star$ and unstructured $v_{ik}^\star$ spatial effects (again representing difference on the spatial structured and unstructured spatial terms between the disease k considered and the reference disease). In our example lung cancer is the reference category ($K = 3$), so that we set $\alpha_3^\star = 0$, $a_{j3}^\star = 0$ (for each age-class $j = 1, \ldots, 13$), $u_{i3}^\star = 0$ and $v_{i3}^\star$ (for each municipalities $i = 1, \ldots, 287$) as constraint for identifiability. Note that terms $u_{i1}^\star$ and $u_{i2}^\star$ represent differences between oral cavity and lung clustering, and between larynx and lung clustering, respectively

$$u_{i1}^\star = u_{i1} - u_{i3} \quad \text{and} \quad u_{i2}^\star = u_{i2} - u_{i3} \quad (2)$$

We consider a model where the difference structured spatial terms (clustering) are decomposed into a shared and a disease-specific effect (Held *et al.*, 2005). We can represent each clustering term for oral cavity and larynx cancer respectively as

$$u_{i1}^\star = ua_i \times \delta_1 + up_{i1} \quad \text{and} \quad u_{i2}^\star = ua_i \times \delta_2 + up_{i2} \quad (3)$$

where ua_i is the shared clustering component (alcohol consumption) and up_{i1} and up_{i2} is the disease specific spatially structured effects. Note that (2) and (3) imply

$$u_{i1} = u_{i3} + ua_i \times \delta_1 + up_{i1} \quad \text{and} \quad u_{i2} = u_{i3} + ua_i \times \delta_2 + up_{i2} \quad (4)$$

Interest is focused on the estimate of disease-specific spatially structured effects because these are considered as latent variables denoting disease-specific risk factors.

It turns out that, after considering known common risk factors, oral cavity and larynx have different specific component spatial patterns that are due to residual risk factors or, more convincingly, to a different interaction between smoking habits and alcohol consumption. In fact, an interaction between smoking and alcohol consumption, suggested by a synergistic effect, has been found repeatedly for oral cavity and larynx cancer, and perhaps we are now able to show that their spatial pattern are slightly different for the two diseases.

References

Dabney A.R. and Wakefield J.C. (2005). Issues in the mapping of two diseases. *Statistical Methods in Medical Research* **14**, 83–112.

Held L., Natário I., Fenton S.E., Rue H. and Becker N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research* **14**, 61–82.

Making sense of large-scale proteomics datasets: myths, facts, and challenges associated with the analysis and statistical evaluation of protein ID and abundance data

Metodi V. Metodiev

1 Introduction

The development of micro-capillary separation methods and new types of hybrid mass spectrometers is revolutionising the field of proteomics. Over the last couple of years advances in instrumentation allowed us to increase more than 10 times the number of proteins that can be measured in a single experiment. However this increased throughput is accompanied with a nearly exponential increase in the size of the raw data files and is already challenging the existing algorithms and software for data analysis.

2 Main results

This talk will discuss the facts and common misconceptions in proteomics data analysis and interpretation in an attempt to compare and contrast the type of information obtained in proteomics and transcriptomics investigations of cancer. The presentation will further discuss data from our ongoing tumour proteomics projects (1, 2) to illustrate the utility of emerging label-free paradigms for protein abundance profiling as tools for the discovery and validation of biomarkers and drug target candidates by direct proteome analysis of fresh and archived tumour tissue specimens.

Metodi V. Metodiev,
Proteomics Unit, Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom e-mail: mmetod@essex.ac.uk

References

Alldridge L., Metodieva G., Greenwood C., Al-Janabi K., Thwaites L., Sauven P., Metodiev M. (2008). Proteome profiling of breast tumors by gel electrophoresis and nanoscale electrospray ionization mass spectrometry. *Journal of Proteome Research*, 7(4), 1458-1469.

Metodieva, G., Greenwood, C., Alldridge, L., Sauven, P., Metodiev, M. (2009). A Peptide-Centric Approach to Breast Cancer Biomarkers Analysis Utilizing Label-Free Multiple Reaction Monitoring Mass Spectrometry. *Proteomics-Clinical Applications*, 3, 78-82.

Specialized Session H

Data Stream Mining

Summarizing and detecting structural drift from multiple data streams

Antonio Balzanella, Rosanna Verde

A growing number of applicative fields is generating huge amount of temporal data. Some examples are rfids, sensors and web logs across industries including manufacturing, financial services and utilities.

In such contexts, data are sequences of values or events obtained through repeated measurements over time. Often these data arrive faster than we are able to mine them so that it may be needed to define new knowledge discovery processes to apply to continuous, high-volume, open-ended data streams.

Algorithms for data streams mining update, in incremental and on-line way, the knowledge about data by means of proper synopses. These provide suitable summaries which are substantially smaller than their base dataset and allow to discard the data just after they have been processed.

Due to the high frequency of data arrival and to the storage constraints imposed by the open ended nature of data streams, often, we have a trade off between accuracy and storage. That is, we generally are willing to settle for approximate rather than exact answers.

An important feature of data streams is their evolution of over time. Whenever it is considered the evolution of the distribution generating the examples of a single data stream, we are dealing with concept drift. When set of streams have to be processed, the temporal evolution implies a change of the proximity relations among the streams, which is usually known as structural drift Gama and Gaber (2007).

Structural drift detection is a very important challenge since it allows to discover global evolutions in the studied phenomenon.

Here, we introduce the basics of a new approach which aims both, at keeping track of the evolution of a set of multiple streams by means of on-line discovered summaries and at measuring such evolution.

Antonio Balzanella,
Second University of Naples, e-mail: antonio.balzanella@gmail.com

Rosanna Verde,
Second University of Naples e-mail: rosanna.verde@unina2.it

Let $S = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ be a set of n streams $Y_i = [(y_1, t_1), \ldots, (y_j, t_j), \ldots, (y_\infty, t_\infty)]$ made by real valued, temporally ordered observations on a discrete time grid $T = \{t_1, ..., t_j, ...t_\infty\} \in \Re$. A time window w_f with $f = 1, \ldots, \infty$ is an ordered subset of T.

The proposal is based on two steps. 1) on-line arriving chunks of data are clustered in a predefined number of clusters and a proximity matrix is updated according to the membership of pairs of sequences to the same cluster. 2) the evolution of proximities among the streams is discovered computing the Frobenius norm of the difference between on-line updated similarity matrices at two time stamps.

The core of the strategy is the on-line update of the proximities among the streams.

The incoming parallel streams are, at first, split into non overlapping windows of fixed size.

A Dynamic Clustering Algorithm (DCA) extended to complex data (Diday, 1971) is, then, run on the subsequences framed by each window w_f in order to get a local partition of the streams in K clusters and a set of prototypes summarizing each cluster.

At each temporal window w_f, starting from the obtained local partition, we keep the proximities among the streams through the squared matrix $\boldsymbol{A}_{\boldsymbol{f}}^{nxn}$. In particular, for each couple of streams Y_i, Y_m allocated to the same cluster, the matrix A_f is updated such that $A_f(i,m) = A_{f-1}(i,m) + 1$, while for each couple of streams allocated to different clusters the set value is $A_f(i,m) = A_{f-1}(i,m)$, where A_{f-1} is the proximity matrix corresponding to the previous window w_{f-1}.

When this procedure is performed on a wide number of windows, it provides an incrementally computed measure of consensus between the couples of streams, obtained taking into account the proximities relations in each local partition.

In order to measure the evolution of the streams, we compute the Frobenius norm $\|\cdot\|_2$ of the matrix $A_f - A_{f-l}$, where l accounts for the desired lag in the evolution monitoring.

References

Diday E. (1971). La methode des Nuees dynamiques, *Revue de Statistique Appliquee*, 19(2), 19-34.

Gama J., Gaber M. M. EDS (2007). *Learning from Data Streams: Processing Techniques in Sensor Networks*, Springer Verlag.

Information-based data stream summary

Fabrice Clérot, Pascal Gouzien

1 Introduction

Data stream mining is becoming ubiquitous because of the increase of the information volumes and the need for real-time access to information and on-line decision-making. Summarizing infinite data streams for fast access to the past information is becoming an active area of research. Such summarization under finite or slowly growing memory usage constraints poses an obvious challenge as a lot of information has to be discarded.

We propose and evaluate a simple window-based scheme which explicitly aims at optimizing this discarding process by maximizing a measure of the information kept in the summary. This is at variance with most classical generic summaries of the literature (Aggarwal et al. (2003), Csernel et al. (2006), see MIDAS (2010) for a survey) where an aging mechanism is deterministically applied: the a priori behind such mechanism is that information is all the more useful as it is more recent. Our summarization scheme borrows from Streamsamp (Csernel et al., 2006) as follows: the time dimension is treated as successive windows of fixed number of events and the summary is organized as a set of such successive (on the time dimension) windows. Under memory usage constraint, as a new window of data has to be inserted into the summary, memory has to be freed by merging two successive windows of the summary.

The merging of two windows of size S results in one window of size S obtained by sampling. Given two successive windows W_1 and W_2, with w_1 and w_2 weights respectively for the sampled elements, the aggregation produces a window W, with a (w_1+w_2) weight. The merging is implemented as a stratified sampling with $S\cdot w_1/(w_1+w_2)$ elements drawn from W_1 and $S\cdot w_2/(w_1+w_2)$ elements drawn from W_2.

Fabrice Clérot, Pascal Gouzien,
Orange Labs Lannion; 2 Avenue Pierre Marzin; 22300 Lannion; France,
e-mail: surname.name@orange-ftgroup.com

2 Main results

Our scheme differs from Streamsamp by the merging rule between successive windows which is biased towards similar windows and does not assume a priori that recent information is more valuable the past information : similarity between windows is measured from the incapacity of a model to discriminate between the windows from the data. The classifier used is Khiops (available as shareware at www.khiops.com) as its performance can be interpreted in terms of information theory: the merged pair is the pair for which the merging minimally increases the entropy of the summary as the repartition of the data between the two windows to be merged is the closest to random noise. Moreover, Khiops allows the treatment of both continuous and categorical data inside the same theoretically consistent framework (Boullé, 2005, 2006).

Experiments comparing this info-based merging rule with random or aging-based merging rules for a constant memory usage constraint show an improvement in the accuracy of the answers to queries on the summary. Further experiments demonstrate how this scheme allows decoupling the memory usage constraint and the merging rule. Moreover, the maintenance of the summary is done in constant time, irrespective of the number of windows. Using a logarithmic memory usage growth constraint, our info-based scheme is compared with Streamsamp.

Finally, because the full Khiops modelling of multivariate data streams can be very time consuming for a large number of variables, we introduce a fast heuristic in the multivariate case and show its efficiency.

All experiments rely on real data taken from a full month of a large web portal audience monitoring at the second level (more than 2.5 million events described as 300-dimensional vectors).

References

Aggarwal C. C., Han J.,Wang J., Yu P. S. (2003). A framework for clustering evolving data streams. In: *VLDB '2003: Proceedings of the 29th international conference on Very large data bases*, 81–92.

Boullé M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes, *Journal of Machine Learning Research*, 6, 1431–1452.

Boullé M. (2006). MODL: a Bayes optimal discretization method for continuous attributes, *Machine Learning*, 65(1), 131–165.

Csernel B., Clrot F., Hbrail G. (2006). Datastream clustering over tilted windows through sampling. *Knowledge Discovery from Data Streams Workshop (ECML/PKDD)*.

MIDAS (an author collective of the French ANR project MIDAS) (2010). Rsum gnraliste de flux de donnes. *EGC '2010*.

Tests for change detection based on incremental quantile estimation

Katharina Tschumitschew, Frank Klawonn

1 Introduction

Quantiles play an important role in statistics, especially in robust statistics, for instance the median as a robust measure of location and the interquartile range as a robust measure of spread. Incremental or recursive techniques for quantile estimation are not as obvious as for statistical moments. Nevertheless, there are techniques for incremental quantile estimation. However, they are either based on a restricted time window (Gelper et al., 2009; Qiu, 1996) or only suitable for continuous random variables. In this paper, we propose a more general approach which is not limited to continuous random variables and we analyse its statistical properties.

For continuous random variables, there is already an incremental scheme for quantile estimation (Nevelson and Chasminsky, 1972; Möller et al. , 2000; Grieszbach and Schack, 1993).

Although this technique of incremental quantile estimation has only minimum memory requirement, it has certain disadvantages.

- It is only suitable for continuous random variables.
- Unless the sequence $\{c_t\}_{t=0,1,\ldots}$ is well chosen, convergence can be extremely slow.
- When the sampled random variable changes over time, especially when the c_t are already close to zero, the incremental estimation of the quantile will remain almost constant and the change will be unnoticed.

In the following we propose a new algorithm to overcome these problems that will also enable us to formulate hypothesis tests for changes of quantiles.

Katharina Tschumitschew,
Ostfalia University of Applied Sciences, Department of Computer Science, Salzdahlumer Str. 46-48, D-38302 Wolfenbüttel, Germany, e-mail: katharina.tschumitschew@ostfalia.de

Frank Klawonn,
Ostfalia University of Applied Sciences, Department of Computer Science, Salzdahlumer Str. 46-48, D-38302 Wolfenbüttel, Germany, e-mail: f.klawonn@ostfalia.de

2 Main results

Let us first consider the online estimation of a specific quantile, the median. Although only one or two exact values (the one or two values in the middle of the ordered data, depending on whether the number of data is odd or even) are needed to calculate the median, it is required to order the data in advance to determine the respective values. Since each time a new data point is added to the data, it is possible for all data points to change their position in the ordered data set. Therefore, in principle all data points must be known or stored for the stepwise computation of the median. The idea of our algorithm is to store only a limited number of exact data values, i.e. values around the median, and to count only the number of data points lying outside an interval around the median. Unfortunately, we do not know the true median and it might turn out that the true median lies outside the interval in which we have stored the exact values. We can, however, compute the probability that this will happen and our algorithm will fail. In this sense, we only provide a probabilistic algorithm which guarantees the correct result only with a certain (very high) probability. This property can also be exploited to formulate hypothesis tests for changes of the median.

Although this idea can be generalised in a straight forward manner to arbitrary quantiles, the performance of such algorithms would not be satisfactory at all. Therefore, we apply a presampling scheme for other quantiles that enables to use the incremental median estimation for the presampled data for the estimation of the quantile of interest.

References

Gelper, S., Schettlinger, K., Croux, C., Gather, U. (2009). Robust online scale estimation in time series: A model-free approach. Journal of Statistical Planning & Inference **139**(2), 335-349

Qiu, G. (1996). An improved recursive median filtering scheme for image processing. IEEE Transactions on Image Processing **5**(4), 646-648

Nevelson, M., Chasminsky, R. (1972). Stochastic approximation and recurrent estimation. Verlag Nauka, Moskau

Möller, E., Grieszbach, G., Schack, B., Witte, H. (2000). Statistical properties and control algorithms of recursive quantile estimators. Biometrical Journal **42**(6), 729-746

Grieszbach, G., Schack, B. (1993). Adaptive quantile estimation and its application in analysis of biological signals. Biometrical journal **35**(2), 166-179

Specialized Session I

Data Visualization

Visualization of clustering comparisons with confusion matrices in Seurat

Alexander Gribov, Antony Unwin

1 Introduction

The analysis of microarray data has become very important in recent years. Medical researchers use them to analyse possible genetic causes for diseases. The reactions of individual genes to testing are described as their expression and these data are displayed in matrix visualizations or heatmaps of genes by patients (Figure 1). Unusually for statistical data, there are far more variables (genes) than cases (patients) and the patients are generally plotted in the columns with the genes in the rows. The visualization for these data involves finding ways of effectively organizing and displaying a matrix of several thousand rows by at most a few hundred columns. One of the strategies developed to deal with it is the use of clustering methods to improve the visualization and analysis of the data. There are many, many different ways of carrying out cluster analyses and it is not easy to decide which one to choose. An important step is to be able to compare two clusterings produced by different approaches. In past a variety of metrics for comparison of clustering results were developed. Unfortunately they are unsuitable to describe complex dependencies and interactions between clusterings. Another way to do this is to draw up a confusion matrix, a table showing how many cases are in both cluster i of the first clustering and cluster j of the second. Careful reordering and visualization of the matrix makes this a powerful tool for comparing clusterings or any classification variables.

Alexander Gribov,
University of Augsburg, Germany e-mail: alexander.gribov@math.uni-augsburg.de

Antony Unwin,
University of Augsburg, Germany e-mail: antony.unwin@math.uni-augsburg.de

2 Main results

The software SEURAT has been developed to provide interactive linked visualizations for all the various kinds of data associated with microarrays. As well as offering heatmaps for gene expression data and standard displays like histograms and barcharts for patient data, it also uses the Rserve package to provide access to the broad range of clustering and seriation tools in R.

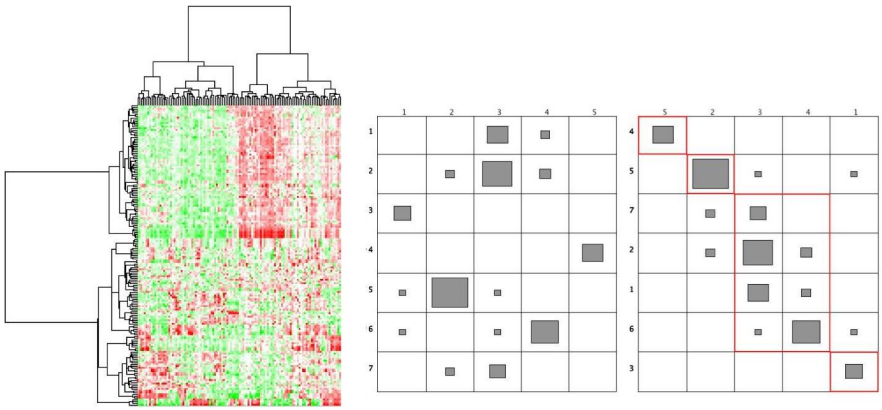


Fig. 1 A heatmap of gene expression data for leukemia patients after clustering (left). Patients in columns of the heatmap were clustered using different methods *ward* and *k-means* in 5 and 7 clusters. The results were compared in SEURAT with confusion matrices: unsorted (center) and sorted (right).

Different clustering results or classifications available in clinical data can be compared using confusion matrices. Displaying the table as a fluctuation diagram in which each table entry is represented by a rectangle whose area is proportional to its count is more effective (cf. the central plot in Figure1). Still more effective is to order the two clusterings in the rows and columns to try to maximize the agreement between them. This problem can be mathematically formulated. A suggested optimization criterion and a heuristic method are described in this paper and have been implemented in SEURAT as an interactive confusion matrix. Linking to all other plots makes it to a powerful tool for visual analysis of clustering results or any other classifications. An example of a sorted confusion matrix in SEURAT is shown in the right-hand plot of Figure 1.

Acknowledgements

This project was supported by the Deutsche José Carreras Leukämie-Stiftung e.V. (Project Number 07/30v).

Visualisation of cluster analysis results

Hans-Joachim Mucha, Hans-Georg Bartel

First, we introduce the problems of finding clusters in a set of objects and of cluster validation. By using special randomized weights of objects one can easily perform built-in validations of cluster analysis results by bootstrapping techniques (Mucha, 2007). Most generally, the stability of cluster analysis results (e.g., hierarchies, partitions, individual clusters, degree of cluster membership of objects) can be assessed based on measures of correspondence between partitions and/or between clusters (Hubert and Arabie, 1985; Hennig, 2007). Here, the finding of the appropriate number of clusters is the main task. However, our proposed built-in validation technique evaluates additionally the stability of each cluster and the degree of membership of each observation to its cluster.

Second, we present some kinds of (multivariate) visualisation of cluster analysis results and of cluster validation results. Visualisation is essential for a better understanding of results because it operates at the interface between statisticians and researchers. Without loss of generality, we will focus on visualisation of clustering based on pairwise distances. Here usually, one can start with "dimensionless" heat plots (fingerprints) of proximity matrices. The Excel 2007 "Big Grid" spreadsheet is both the distinguished repository for data/proximities and the perfect plotting board for (multivariate) graphics such as dendrograms, plot-dendrograms, informative dendrograms, scatterplot matrices, density plots, principal components analysis plots and discriminant projection plots (Mucha, 2009). Informative dendrograms are ordered binary trees (Mucha and Bartel and Dolata, 2005) that show additional information such as stability values or other descriptive statistics. The programming language is Visual Basic for Application (VBA). Really, in Excel 2010, you get much more because it offers new kinds of built-in visualisations. For exam-

Hans-Joachim Mucha,
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), D-10117 Berlin, Germany,
e-mail: mucha@wias-berlin.de

Hans-Georg Bartel,
Institute for Chemistry, Humboldt University Berlin, Brook-Taylor-Straße 2, D-12489 Berlin, Germany, e-mail: hg.bartel@yahoo.de

ple, sparklines (small cell-sized graphics) play an important (additional) role for an improved visualisation of cluster analysis results.

References

Hennig C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52, 258–271.

Hubert L. J., Arabie P. (1985). Comparing Partitions. *Journal of Classification*, 2, 193–218.

Mucha H.-J. (2007). On Validation of Hierarchical Clustering. In: R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis*. Springer, Berlin, 115–122.

Mucha H.-J. (2009). Cluscorr98 for Excel 2007: Clustering, Multivariate Visualization, and Validation. In: H.-J. Mucha and G. Ritter, editors, *Classification and Clustering: Models, Software and Applications*, *Report* no. 26, WIAS, Berlin, 14–41.

Mucha H.-J., Bartel H.-G., Dolata J. (2005). Techniques of Rearrangements in Binary Trees (Dendrograms) and Applications. *Match* 54(3), 561–582.

Some recent advances on dimension reduction for high-dimensional data

Luca Scrucca

1 Introduction

High dimensional data represent a challenging task for many statistical procedures. An extreme form of high-dimensional data is the so-called "large p small n" case, i.e. the situation $n \ll p$ where n denotes the sample size and p the number of predictors in a regression problem. This type of data often arise in various disciplines, such as chemistry, medicine and biology. Partial least squares (PLS) is a well-known methodology for studying the regression of a response variable Y on a set of p predictors X when p is large. PLS has now become a standard tool for regression modeling in chemometrics.

Dimension reduction (DR) methods in regression aims at reducing the predictor dimension while preserving full information for the regression of Y on X, and without assuming any parametric model. Thus, they are particularly valuable when p is large and the predictors are highly correlated. One of the most popular DR method is sliced inverse regression (SIR; Li, 1991).

Recently, Li et al. (2007), borrowing ideas from the relationship of PLS with ordinary least squares (OLS), extended the application of SIR to the case $n \ll p$. In this paper we investigate the extension of Principal Fitted Components (PFC; Cook, 2007; Cook and Forzani, 2008) to the case of high dimensional data.

2 Main results

The main objective of DR methods is the estimation of the basis $\beta \in \mathbb{R}^{p \times d}$ of the central subspace $\mathscr{S}_{Y|X}$, with $d \leq p$, such that $Y \perp\!\!\!\perp X | \beta^\top X$. The spanning matrix β can be estimated using several methods, such as OLS, SIR, SAVE, pHd, etc.

Luca Scrucca,
Dipartimento di Economia Finanza e Statistica, Università degli Studi di Perugia,
e-mail: luca@stat.unipg.it

PLS is a method for estimating $q \leq p$ latent variables such that they are most correlated with the response variable. Following Helland (1988), the PLS coefficients vector for q latent factors can be expressed as $\beta_{\text{PLS}} = \text{P}_{R_q(\Sigma_X)}\beta_{\text{OLS}}$, where $\beta_{\text{OLS}} = \Sigma_X^{-1}\sigma_{XY}$ is the OLS coefficients vector, $\text{P}_{R_q(\Sigma_X)}$ denotes the projection onto the subspace spanned by the columns of the Kyrlov sequence $R_q = (\sigma_{XY}, \Sigma_X\sigma_{XY}, \ldots, \Sigma_X^{q-1}\sigma_{XY})$ with respect to the Σ_X inner product.

Li et al. (2007) used this result to introduce the partial inverse regression estimator (PIRE) given by $\beta_{\text{PIRE}} = \text{P}_{R_q(\Sigma_X)}\beta_{\text{SIR}}$, where $\beta_{\text{SIR}} = \Sigma_X^{-1}\zeta_{X|Y}$ with $\zeta_{X|Y}$ being the first principal eigenvector of $\Sigma_{X|Y} = \text{Var}(\text{E}(X|Y))$, and $R_q = (\zeta_{X|Y}, \Sigma_X\zeta_{X|Y}, \ldots, \Sigma_X^{q-1}\zeta_{X|Y})$. PIRE is thus an extension of SIR which can be used also in the case of $n \ll p$. The authors showed that it can be effectively employed in those cases where the mean function $\text{E}(Y|X)$ is curved, or the dependence is through the variance function $\text{Var}(Y|X)$. However, in cases where the mean function is linear the performance of PLS is superior.

In this paper we investigate the extension of PFC to the case $n \ll p$ by using the above approach based on Krylov sequence. Let $\beta_{\text{PFC}} = \Sigma_X^{-1}\nu_{X|Y}$, where $\nu_{X|Y}$ is the first principal eigenvector of Σ_{fit}, the covariance matrix of the fitted vectors from the multivariate linear regression of X on F_Y. This last term can be any transformation of the response Y, and often a cubic polynomial is used. PFC can be seen as a different way of approximating the inverse mean function $\text{E}(X|Y)$; in contrast, SIR approximates each conditional inverse mean $\text{E}(X_j|Y)$ by a step function of Y with h steps, where h is the number of slices used to partition the range of Y. The partial principal fitted components (PPFC) estimator can be computed for q latent factors as $\beta_{\text{PPFC}} = \text{P}_{R_q(\Sigma_X)}\beta_{\text{PFC}}$, where $\text{P}_{R_q(\Sigma_X)}$ denotes the projection onto the subspace spanned by the columns of the Kyrlov sequence $R_q = (\nu_{X|Y}, \Sigma_X\nu_{X|Y}, \ldots, \Sigma_X^{q-1}\nu_{X|Y})$ with respect to the Σ_X inner product. We found that the PPFC estimator (i) is comparable to PLS when the regression mean function $\text{E}(Y|X)$ is linear, (ii) appears to improve over the PIRE estimator when the mean function is nonlinear or the dependence is in the variance function.

References

Cook D.R. (2007) Fisher lecture: dimension reduction in regression. *Statistical Science*, 22 (1), 1–26.

Cook D.R., Forzani L. (2008) Principal fitted components for dimension reduction in regression. *Statistical Science*, 23 (4), 485–501.

Helland I.S (1988) On the structure of partial least squares regression, *Communication in Statistics: Simulation and Computation*, 17, 581-607.

Li K.C. (1991) Sliced inverse regression for dimension reduction (with discussion) *Journal of the American Statistical Association*, 86, 316–342.

Li L., Cook D.R., Tsai C.L. (2007) Partial inverse regression. *Biometrika*, 94 (3), 615–625.

Specialized Session J

Classification in Systems Biology II

Phylogenomics as a standard technique for microbial taxonomy

Markus Göker

Since the pioneering work of Carl Woese and his coworkers in the late Seventies, three main branches of the tree of life have been established: Bacteria, Archaea, and Eukaryotes. While the Prokaryotes (Bacteria and Archaea) belong to the most inconspicuous living beings on earth, they by far outclass the Eukaryotes regarding their biochemical abilities, and most likely also regarding the total number of species and the total biomass. Because of their small size and the frequent lack of morphological distinctions, the classification of Prokaryotes is dominated by chemotaxonomical and molecular techniques and has been considerably standardized. Due to the recent staggering advances in DNA sequencing technology, Prokaryotic genomes can be obtained in steadily decreasing time and at steadily decreasing costs. Given the considerable promise whole-genome sequencing offers for phylogeny and classification, it is surprising that microbial systematics and genomics have not yet been reconciled. This might be either due to the intrinsic difficulties in inferring reasonable phylogenies from genomic sequences, particularly in the light of the significant amount of lateral gene transfer in prokaryotic genomes, or simply caused by the current lack of completely sequenced genomes for many of the major lineages of prokaryotes and for most type strains. Here, I argue that the species tree and the hierarchical classification that is based on it are still meaningful concepts, and that state-of-the-art phylogenetic inference methods are able to provide reliable estimates of the species tree to the benefit of taxonomy. Thus, phylogeny-driven microbial genome sequencing projects such as the Genomic Encyclopaedia of *Archaea* and *Bacteria* (GEBA) project are likely to rectify the current situation. However, establishing a fully genome-based classification system will present considerable computational challenges.

Markus Göker
DSMZ, Braunschweig, Germany, e-mail: markus.goeker@dsmz.de

Boolean networks for modeling gene regulation

Hans A. Kestler

1 Introduction

At the core of systems biology research lies the identification of biomolecular networks from experimental data via reverse-engineering methods. In this context, Boolean networks provide a framework for reverse-engineering and analysis of gene-regulatory networks (Kauffman, 1969; Müssel et al., 2010). In a Boolean network, a gene is modeled as a Boolean variable over discrete time that can attain two alternative levels, expressed (1) or not expressed (0). Liang et al. (1998) developed the REVEAL algorithm, which uses the mutual information between input and output states (e.g., two subsequent measurements of a time series) to infer Boolean dependencies between them. Akutsu et al. (2000) presented an algorithm that can deal with noisy time series data to infer such networks. Lähdesmäki et al. (2003) derive a method that reduces the time complexity of this search (by a factor of 2^{2^k} compared to Akutsu et al. (2000) where k is the number of input genes of a single Boolean function). A general problem in reconstructing networks from time series data is the large number of different genes compared to a relatively low number of temporal measurement points. As a result, the available data are often consistent with multiple network configurations. A further problem is, that high-throughput techniques like microarray analyses provide real-valued profiles, hence the data has to be binarized carefully. However, the noisiness of gene expression data and the low number of temporal measurement points often yield several plausible binarizations. Differences in the binarization results can have strong effects on the architecture of the resulting Boolean networks because a state change for a single gene can lead to many differences in downstream functions and gene dependencies.

H. A. Kestler
Institute of Neural Information Processing, University of Ulm, 89069 Ulm, Germany
Department of Internal Medicine I, University Hospital Ulm, 89081 Ulm, Germany
Tel.: +49-731-5024248
Fax: +49-731-5024156
e-mail: hans.kestler@uni-ulm.de

2 Main results

To address some of these issues, two scale-space binarization method were developed to produce suitable and robust thresholds even for small numbers of data points. Additionally a measure of validity for the found thresholds is given. Incorporating this knowledge into network reconstruction allows for restricting the input of reconstruction algorithms to genes with meaningful thresholds. This reduces the complexity of network inference. The performance of our binarization algorithms was evaluated in network reconstruction experiments using artificial data as well as real-world yeast expression time series. The new approaches yield considerably improved correct network identification rates compared to other binarization techniques by effectively reducing the amount of candidate networks.

References

Akutsu T., Miyano S., Kuhara S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8), 727-734.

Kauffman S. A. (1969). Metabolic Stability and Epigensis in Randomly Constructed Genetic Nets. *Journal of Theoretical Biology*, 22, 437-467.

Lähdesmäki H., Shmulevich I., Olli Yli-Harja O. (2003). On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning*, 52(1-2), 147–167.

Liang S., Fuhrman S., Somogyi R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In: Altman R. B., Dunker A. K., Hunter L., Klein T. E. D. EDS, Proceedings of the Pacific Symposium on Biocomputing, World Scientific, Vol.3, 18-29.

Müssel C., Hopfensitz M., Kestler H. A. (2010). BoolNet-an R package for generation, reconstruction, and analysis of Boolean networks. *Bioinformatics*, 26(10), 1378-1380.

Graphical models for eliciting structural information

Federico M. Stefanini

1 Introduction

Bayesian Networks (BN) are increasingly exploited to represent probabilistic and causal relationships in biology, for example in the so called 'omics' fields (Wilkinson, 2007). Learning the structure of a BN is still challenging for the combinatorial explosion of candidate structures with the increase of the number of nodes.

Bayesian structural learning of BNs depends on the joint distribution function of data, parameters and structure, or the marginal distribution after integrating out model parameters. Widely adopted elicitation techniques define the initial distribution on the space of Directed Acyclic Graphs (DAGs) Ω_Z given a fixed set V of nodes by placing restrictions like: a total ordering of nodes, the presence of sharp order constraints on some nodes, the marginal independence of unknowns or the existence of a prior network which is a good summary of expert prior beliefs. All these (and others) restrictions aim at making the structural learning task feasible even in medium-to-large networks.

Trascriptomics, proteomics, metabolomics and other 'omics' fields at the core of system biology suffer the '$n << p$' curse of dimensionality, thus problem regularization is often invoked through a prior distribution on the space of DAGs. In these fields expert information may be absolutely relevant although focused on just some limited aspects of the problem domain, for example a subset of nodes-genes, an hypothetical causal path of gene action or the existence of a node-protein which acts as a regulatory hub. In a general setting the available prior information may involve 'network features' which are only indirectly related to each arrow of the network, therefore an approach to elicitation suited to system biology should put reference network features (Stefanini, 2008) as preeminent building blocks of the elicitation procedure.

Federico M. Stefanini,
Department of Statistics 'G.Parenti', University of Florence, Italy e-mail: stefanini@ds.unifi.it

Here we generalize recent proposals from the literature (Stefanini, 2009a,b) by introducing an approach based on chain graph models which represent the expert prior belief about relevant network features. Seminal contributions to structural elicitation from the literature are casted in the proposed framework to reveal their strength, limitations and to disclose natural generalizations allowing more expressiveness. We finally present some auxiliary procedures to tune and revise the elicited prior distribution.

2 Main results

Let $R = (R_1, R_2, \ldots, R_{n_r})$ be a vector associated with the presence/absence of n_r reference network features. The main equation to elicit the probability of a structure $z \in \Omega_Z$, given the context information ξ, is:

$$P[Z = z \mid \xi] = \frac{1}{n_z} \cdot P[R = (r_1, r_2, \ldots, r_{n_r}) \mid \xi] \tag{1}$$

where n_z is the cardinality of the equivalence class of DAGs in which z is located together with all DAGs characterized by the same configuration of features as coded by R. Providing a Monte Carlo estimate of n_z, the probability in the r.h.s. of (1) is the degree of belief that the unknown structure has a configuration of reference features as coded by vector R. We show that the distribution $p_R(r_1, r_2, \ldots, r_{n_r} \mid \xi)$ may be usefully coded to be Markovian with respect to a chain graph. A partial order relation elicited on the set of reference features defines subsets of nodes in each chain component τ_g. The elicitation of probability distributions associated to undirected graphs defined in each chain component is performed using suitable odds against the 'no feature' configuration taken as reference.

The proposed approach is illustrated by reconsidering previous seminal proposals from the literature in a unified and formalized setup.

References

Stefanini F. M. (2008). Eliciting expert beliefs on the structure of a Bayesian Network, PGM2008 - Probabilistic Graphical Models 2008, Hirtshals, Denmark.

Stefanini F. M. (2009a). The revision of elicited beliefs on the structure of a Bayesian Network, S.Co. 2009, Milano, Italy.

Stefanini F. M. (2009b). Prior beliefs about the structure of a probabilistic network, SIS2009, Pescara, Italy.

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, 8 (2), 109-116.

Specialized Session K

Social Networks and Classification

Classification of personal networks using latent roles

Ulrik Brandes, Jürgen Lerner, Miranda J. Lubbers, Christopher McCarty, José Luis Molina, Uwe Nagel

1 Introduction

We present a method for graph classification based on the assumption that members of the same class have a similar role structure and that these roles can be derived from a joint classification of all vertices (Brandes et al., 2010). Note that this a different from graph clustering, blockmodeling, or role assignment, where one is interested in partitioning the vertices of a single graph.

Our approach is motivated by social network concepts, and we demonstrate its utility on an ensemble of personal networks of migrants. Personal networks are studied because, in addition to personal attributes such as age, gender, race, job position, or income, the composition and structure of a persons's social network can be indicative of inter-personal variance (McCarty, 2002).

2 Main results

Assume we are given an ensemble of networks, where a network is a graph together with vertex attributes, and an ensemble is a set of networks of the same type with a common origin. The proposed method for classification consists of three main steps:

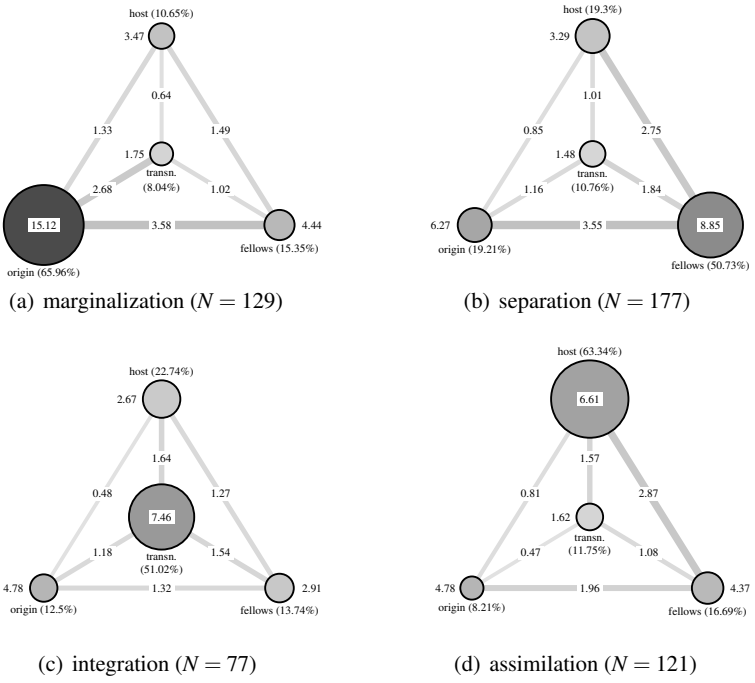
1. Determine node classes by jointly clustering the vertices of all networks based on their attribute vectors.

Ulrik Brandes, Jürgen Lerner, Uwe Nagel
Computer & Information Science, University of Konstanz, e-mail: ulrik.brandes@uni-konstanz.de

Miranda J. Lubbers, José Luis Molina,
Social & Cultural Anthropology, Autonomous University Barcelona,

Christopher McCarty,
Bureau of Economics and Business Research, University of Florida

2. Project each graph onto the common role graph, i.e, contract vertices of the same class into a class node.
3. Derive feature vectors from the resulting role graphs and cluster them using any suitable method.



(a) marginalization ($N = 129$) (b) separation ($N = 177$)

(c) integration ($N = 77$) (d) assimilation ($N = 121$)

Fig. 1 In our application example, we use the immigration status for node classes and find four clusters of personal networks as summarized above. These can actually be labeled with the modes of acculturation from Berry (1997).

References

Berry J. W. (1997). Immigration, acculturation, and adaptation. *Applied Psychology*, 46(1), 5–68.

Brandes U., Lerner J., Nagel U. (2010). Classification in network ensembles using latent roles. Submitted for publication in *Advances in Data Analysis and Classification*.

McCarty C. (2002). Structure in personal networks. *Journal of Social Structure*, 3(1).

Developments in blockmodeling

Anuška Ferligoj

One of the major goals of social network analysis is to discern *fundamental structures* of networks in ways that allow us to get insight into the structure of a network and to facilitate our understanding of network phenomena by providing a set of tools that includes ways of mapping social structures in ways that help identify positions and roles. The key tasks here are:

- identifying *social positions* as collections of units who are similar in their relationships to the others, and
- modeling *social roles* as system of relationships among positions (Faust and Wasserman, 1992).

Blockmodeling is dealing with these two aspects. The goal of *blockmodeling* is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped into the positions according to the extent to which they are equivalent, according to some *meaningful* definition of equivalence.

Lorrain and White (1971) introduced the concept of structural equivalence: units are equivalent if they are connected to the rest of the network in identical ways. Sailer (1978) provided another way of thinking about blockmodeling. This was later formalized by White and Reitz (1983) with the introduction of regular equivalence as a formal generalization of structural equivalence. Units are regular equivalent if they link in equivalent ways to other units that are also regular equivalent. In 1992, the journal Social Networks devoted a special issue to blockmodeling featuring a variety of approaches that had been created since the early statements. This helped to create the conditions for the emergence of generalized blockmodeling as a systematic statement of the blockmodeling approach (Doreian et al., 2005).

In the presentation recent extensions to generalized blockmodeling and some open blockmodeling problems will be discussed.

Anuška Ferligoj,
University of Ljubljana, e-mail: anuska.ferligoj@ffdv.uni-lj.si

The following extensions to generalized blockmodeling recently appeared:

- two-mode and three mode network arrays,
- valued networks,
- stochastic blockmodeling,
- generalized blockmodeling of signed networks,
- partitioning large and/or complex networks.

There are still several open generalized blockmodeling problems. The following ones will be discussed:

- boundary problems,
- measurement errors,
- assessing fits of blockmodels,
- blockmodeling large networks,
- numbers of positions,
- dynamic blockmodels.

References

Doreian P., Batagelj V., Ferligoj A. (2005). *Generalized Blockmodeling*, Cambridge: Cambridge University Press.

Faust K. and Wasserman S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks*, 14, 5-61.

Lorrain F. and White H.C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1, 49-80.

Sailer L.D. (1978). Structural equivalence: Meaning and definition, computation and application. *Social Networks*, 1, 73-90.

White D.R. and Reitz K.P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5, 193-234.

Clustering social actors by using auxiliary information on relational data

Giuseppe Giordano, Maria Prosperina Vitale

1 Introduction

The aim of this paper is to show how different kinds of auxiliary information can be used to describe relational data structures and reveal hidden patterns of homogeneous social actors behaviors. We consider an affiliation matrix and two sets of variables describing the characteristics of nodes (actors) and events. The two auxiliary data structures will be exploited to take into account the relationships between categories of nodes/events and their influence on relational data.

The network data, considered as a set of contiguity constraints on statistical units, has been regarded in Multidimensional Data Analysis techniques to define local factorial analysis (Aluja, 1984). On the other hand, in Social Network Analysis the attribute data are mainly used to describe homogeneous groups of actors or to explain the probability of forming ties in statistical relational models (Wasserman, 1994). We study how common characteristics observed on both actors and events in a two-mode network can affect the participation of actors to events. The underlying theory, known as *Homophily effect* (McPherson, 2001), stems from the principle that contacts between similar people occur at higher rates than among dissimilar people.

We introduce a data analysis strategy based on some derived network structures which can highlight the presence of peculiar patterns. Our leading hypothesis is that the role and position of actors in social networks can be studied according to two perspectives: the reciprocal influence of network and auxiliary data and the definition of the metric measurement of ties. In particular, the obtained results make it possible to represent homogeneous groups of social actors described by auxiliary information.

Giuseppe Giordano,
Department of Economics and Statistics, University of Salerno, e-mail: ggiordan@unisa.it

Maria Prosperina Vitale,
Department of Economics and Statistics, University of Salerno, e-mail: mvitale@unisa.it

2 Main results

Let $\mathbf{A}$ $(n \times p)$ be the affiliation matrix, where n is the number of nodes and p is the number of events, and the entries represent the frequency of participation of the actors to the events. Let $\mathbf{X}$ $(n \times m)$ be the matrix holding the auxiliary information on nodes, where m is the number of auxiliary variables (either metric or non-metric). Let $\mathbf{Z}$ $(q \times p)$ be the matrix holding auxiliary information collected through q characteristics observed on the p events. In the following, we consider the case of complete disjunctive codings of nominal auxiliary variables - in $\mathbf{X}$ and $\mathbf{Z}$ - and a binary relation in the matrix $\mathbf{A}$.

We set the matrix $\mathbf{B}$ $(m \times p) = [diag(\mathbf{X}'\mathbf{X})]^{-1}\mathbf{X}'\mathbf{A}$ in which the elements can describe the participation of nodes category to the events, and the matrix $\mathbf{C}$ $(q \times n)$ $= [diag(\mathbf{Z}\mathbf{Z}')]^{-1}\mathbf{Z}\mathbf{A}'$, holding the relative frequencies of participation of actors to kinds of event.

Starting from the above defined matrices, $\mathbf{B}$ and $\mathbf{C}$, we may derive corresponding valued adjacency matrices, $\mathbf{G}_X$ (1) and $\mathbf{G}_Z$ (2).

$$\mathbf{G}_X = \mathbf{X}\mathbf{B}\mathbf{B}'\mathbf{X}'; \tag{1}$$

$$\mathbf{G}_Z = \mathbf{C}'\mathbf{Z}\mathbf{Z}'\mathbf{C}; \tag{2}$$

The analysis of $\mathbf{G}_X$ and $\mathbf{G}_Z$ can carried out through two steps. In the first step we look for homogeneous patterns in the adjacency data. At this aim, the nodes are clustered according to hierarchical classification algorithm. The partition induced by the classification is used to re-arrange the rows and columns of the adjacency matrix in order to visualise the permuted matrix. In the second step, once obtained a suitable partition, each group of nodes is described according to their characteristics. Each group can be represented as a super-node described by logical operators joining the weights or frequencies of the auxiliary variables. The proposed approach allows to enrich the traditional network analysis by different point of view: i) the complementary use of valued graphs defined according to observed auxiliary information; ii) the possibility to introduce explicative measures linking auxiliary information and relational data; iii) the interpretation of the analysis results as complex data where classes of social actors are defined and interpreted as second-order individuals.

References

Aluja Banet T., Lebart L. (1984). Local and Partial Principal Component Analysis and Correspondence Analysis. In: Havranek, T., Sidak, Z., Novak, M. (Eds.) COMPSTAT Proceedings, Phisyca-Verlag: Vienna, 113-118.

McPherson M., Smith-Lovin L., Cook. J. (2001). Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology *27*, 415-444.

Wasserman S., Faust K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Specialized Session L

Statistical Matching: Theory and Applications to Data Mining and Official Statistics

Matching two different topics: ecological inference and data fusion

Marcello D'Orazio, Marco Di Zio, Mauro Scanu

1 Introduction and objective of the research

Sometimes different topics deal with the same problem and are tackled with different tools. This is the case of data fusion (also known as statistical matching) and ecological inference. Data fusion techniques (D'Orazio *et al.*, 2006) are aimed to combine information available in distinct sources of data referred to the same target population. In the usual framework we generally need to integrate two data sets A and B that contain data collected in two independent sample surveys such that (i) the two samples contain distinct units (the samples do not overlap); (ii) the two samples contain information on some variables X (common variables), while other variables are observed distinctly in one of the two samples, say, Y in A and Z in B. The goal is the estimation of the relationship between Y and Z by exploiting the knowledge of the common variables X. Unless external information is used, it is not possible to identify a unique model because the lack of joint information on the variables Y and Z. One of the most common external information is the assumption of conditional independence between Y and Z given X. Strategies for alleviating conditional independence assumption (CIA) are based on the use of auxiliary information concerning Y and Z, coming for instance from an outdated survey. When auxiliary information is not available or when we do not want to use it because it is not considered reliable, the analysis of 'uncertainty', as suggested by D'Orazio *et al.* (2006) (as well as in other older papers and books cited therein, as the ones by Kadane, Moriarity and Scheuren, and Raessler), is useful. It consists on the analysis of all the models compatible with the data at hand. According to this approach the estimate of the parameters are not anymore given by a single value but by an interval. It is important to remark that this interval is conceptually different from the inference based on confidence intervals where the uncertainty taken into account is

Marcello D'Orazio, Marco Di Zio, Mauro Scanu,
Istat, e-mail: madorazi@istat.it, e-mail: dizio@istat.it, e-mail: scanu@istat.it

due to sampling variability, while in this case the uncertainty is due to the lack of information that implies model unindentifiability.

Ecological inference is essentially an inferential problem with partial information. Missing information consists of the relationship between the variables of interest. The first approaches focused on the definition of those models that allow the problem to be identifiable, i.e. estimable, according to the data at hand. A prominent example of ecological inference relates to the evaluation of the voting behaviour in political elections (King, 1997) when there is only partial information in terms of: the proportion of voting-age population who are black (p_i), the proportion of voting-age population turning out to vote (q_i), the number of people of voting-age (N_i). These data are available for all the precincts in a certain geographic area. The aim is to estimate the proportion of voting-age blacks who vote (β_{bi}) and the proportion of voting age whites who vote (β_{wi}) for each precinct i. A fundamental equation in ecological inference is synthesized by the ecological regression equation, known also as tomography line:

$$q_i = \beta_{bi} p_i + \beta_{wi}(1 - p_i). \tag{1}$$

Usually this model is simplified and made estimable assuming that the probabilities $\beta_{bi} = \beta_b$ and $\beta_{wi} = \beta_w$ are the same in the different precincts. As a matter of fact, this model is a different conditional independence assumption than the one usually used in statistical matching problems. The traditional conditional independence assumption of Y and Z given X is assumed in Freedman *et al.* (1991). As in statistical matching, the identifiable models have been largely criticized in ecological inference, see Chambers and Steel (2001), King (1997) and Wakefield (2004).

References

Chambers, R.L., Steel, D. G. (2001). Simple methods for ecological inference in 2x2 tables. *Journal of the Royal Statistical Society*, A, 164, 175-192

D'Orazio M., Di Zio M., Scanu M. (2006). *Statistical Matching: Theory and Practice*, Chichester: Wiley.

Freedman, D., Klein, S., Sacks, J., Smyth, C.A., Everett, C.G. (1991). Ecological regression and voting rights. *Evaln. Rev.*, 15, 673-711.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Wakefield, J. (2004). Ecological inference for 2x2 tables. *Journal of the Royal Statistical Society*, A, 167, 385-445.

Displaying uncertainty in data fusion by imputation

Florian Meinfelder, Susanne Rässler

Data fusion techniques typically aim to achieve a complete data file from different sources which do not contain the same units. Traditionally, data fusion, in the US also addressed by the term statistical matching, is done on the basis of variables common to all files. It is well known that those approaches establish conditional independence of the (specific) variables not jointly observed given the common variables, although they may be conditionally dependent in reality. However, if the common variables are (carefully) chosen in a way that already establishes conditional independence, then inference about the actually unobserved association is valid. In terms of regression analysis, this implies that the explanatory power of the common variables is high concerning the specific variables. Unfortunately, this assumption is not testable yet. Hence, we treat the data fusion situation as a problem of missing data by design and suggest imputation approaches to multiply impute the specific variables using informative prior information to account for violations of the conditional independence assumption. In a simulation study it is also shown that multiple imputation techniques allows to efficiently and easily use auxiliary information.

Florian Meinfelder, Susanne Rässler,
University of Bamberg, Germany

Data fusion for direct marketing

Peter van der Putten

1 Introduction

With no data, there is nothing to mine in. Multiple of sources of data can exist, and linking this data together can be non trivial. Assume we are given an instance, representing for example a customer. The problem of merging information from different sources about this particular instance, assuming it can't be done with simple joins, is also called the exact matching problem (Radner et al. (1980)). In contrast, enriching the data for this instance with information from other instances is called a statistical matching or data fusion problem, which is the topic of our research.

In literature data fusion is almost exclusively used in a market research or socio-economic survey context, to merge information from samples with different sets of questions, to reduce the response burden or to connect survey data that has previously not been studied jointly. The resulting surveys are then typically mined using simple techniques such as cross tabulations and correlation analysis. However for direct marketing, information is required for every single customer to allow for personalized one to one communication, so typically a customer database is fused with surveys. The fused data can then directly be exploited, for instance for rule based segmentation (van Hattum and Hoijtink (2008)), or the enriched data can be used to build prediction models with an improved performance or that are easier to understand (van der Putten et al. (2002)).

The goal of our paper is not to present new (or any) algorithms, but it is an application oriented position paper to discuss opportunities and challenges for applying data fusion in a data mining for direct marketing context. To guide the discussion we will use a a proof of principle example that demonstrates data fusion can add value for predictive modeling for direct marketing (van der Putten et al. (2002); van der Putten (2010)). The results will generalize to any case where there is an interest to enrich data that will then be used for predictive modeling.

Peter van der Putten,
LIACS, Leiden University, The Netherlands, e-mail: putten@liacs.nl

Table 1 External evaluation results: using enriched data generally leads to improved performance in this example.

	Only common variables	**Common and correlated fusion variables**	**Common and all fusion variables**
SCG neural network	c=0.692 $\pm$ 0.012	c=0.703 $\pm$ 0.015 p=0.041	c=0.694 $\pm$ 0.019 p=0.38
Linear regression	c=0.692 $\pm$ 0.014	c=0.724 $\pm$ 0.012 p=0.000	c=0.713 $\pm$ 0.013 p=0.002
Naive Bayes Gaussian	c=0.701 $\pm$ 0.015	c=0.720 $\pm$ 0.012 p=0.003	c=0.719 $\pm$ 0.012 p=0.005
Naive Bayes multinomial	c=0.707 $\pm$ 0.015	c=0.720 $\pm$ 0.011 p=0.200	c=0.704 $\pm$ 0.009 p not relevant
***k*-nearest neighbor**	c=0.702 $\pm$ 0.012	c=0.716 $\pm$ 0.013 p=0.0093	c=0.720 $\pm$ 0.012 p=0.0023

2 Main results

In Table 1 the results of the proof of principle experiment can be found. Using real world data we simulated a use case in which a customer database was enriched with survey data through data fusion, and models were built to predict credit card ownership with or without fused survey data using a variety of algorithms. Using the survey data generally leads to better models (the *c* measure is similar to AUC). This example can then be used to discuss challenges and opportunities for applying data fusion for direct marketing.

References

Radner D. B., Rich A., Gonzalez M. E., Jabine T. B., Muller H. J. (1980). Report on Exact and Statistical Matching Techniques, Statistical Working Paper 5, Technical Report, Office of Federal Statistical Policy and Standards US DoC.

van der Putten, P. (2010), On Data Mining in Context: Cases, Fusion and Evaluation, PhD thesis, Leiden Institute of Advanced Computer Science (LIACS), Leiden University.

van der Putten P., Kok J. N., Gupta A. (2002). Why the information explosion can be bad for data mining, and how data fusion provides a way out, in: R. L. Grossman, J. Han, V. Kumar, H. Mannila, R. Motwani EDS, *Second SIAM International Conference on Data Mining (SDM 2002)*, SIAM, 128–138.

van Hattum P., Hoijtink H. (2008). The Proof of the Pudding is in the Eating. Data Fusion: an Application in Marketing, *Journal of Database Marketing and Customer Strategy Management*, 15(4), 267–284.

Contributed Session 1

Achievement and Research

Bibliometric indicators for statisticians: critical assessment in the Italian context

Francesca De Battisti, Silvia Salini

1 Introduction

The evaluation of the university and scientific research has become increasingly important in recent years. In particular, there is a growing interest in the evaluation of scientific publications and related bibliometric indicators (Marchant, 2009). The new criteria acquired in the university context, setting up the funding on the basis of assessments of the scientific productivity of universities and departments, as well as regulating the career advancement of individuals assessing their research products, require careful examination of databases available in different fields and kinds of information obtained from their query. It is important to notice that bibliometric indicators can not be self-sufficient instruments of assessment, but they must be integrated into more complex system of assessment; their oversimplified use, oriented to reduce the complexity of the evaluation, would have a severely negative impact on the resulting decision-making process. Despite that, the output of the databases is the image that the international reviewers (of journals, research projects, visiting demands and partnerships) have about the Italian statistics researchers and scientific community. Knowing of operational limitations about use, coverage and updating of databases (Falagas et al, 2008), the aim of this research is to gain awareness and knowledge of the image, true or false, obtained by them: the study analyses the scientific production of all italian statistics academic scholars (SECS/S01).

2 Main results

The databases that will be considered are:

1. Current Index to Statistics (CIS), created by the American Statistical Association and the Institute of Mathematical Statistics (http://www.statindex.org/).

Francesca De Battisti,
University of Milan e-mail: francesca.debattisti@unimi.it

Silvia Salini,
University of Milan e-mail: silvia.salini@unimi.it

2. Web of Science (WoS), edited by the Institute for Scientific Information and distributed by Thomson Reuters (http://isiwebofknowledge.com/).
3. Scopus, sponsored by Elsevier (www.info.scopus.com).
4. Google Scholar, with recommended interface Publish or Perish, developed by Anne-Wil Harzing (http://www.harzing.com/pop.htm).

By the database query, made in the period from February to April 2010, a dataset was built, in which there are the variables: number of publications for each database, corresponding time period and, excluding CIS, number of citations and h-index (Marchant, 2009). There are also descriptive variables such as title and affiliation, obtained by MIUR. Table 1 shows the joint distribution of the number of publications of italian researchers according to the CIS and WoS databases.

Table 1 Number of publications on CIS vs Number of publications WoS

		WoS						Total
		<= 5	6 - 10	11 - 15	16 - 20	21 - 25	26+	
CIS	<= 5	203	21	2	0	0	0	226
	6 - 10	71	23	5	1	1	0	101
	11 - 15	24	18	10	1	0	0	53
	16 - 20	2	8	5	5	2	1	23
	21 - 25	5	7	1	1	1	1	16
	26+	6	1	6	4	4	4	25
Total		311	78	29	12	8	6	444

First of all, the SECS/S01 scholars will be classified on the basis of 10 quantitative variables obtained from the databases, adding an additional dichotomous variable for each person that points out whether or not the subject has published on the "top five" journals resulting from the SIS Survey[1]. A preliminary classification shows that there is a group of "better" researchers, that have high values on all variables, a group of scholars who publish much but have less citations, others have a lot of papers in other fields than statistics, etc. As a second step, using data reduction techniques, latent variables that give reason for the detected clusters, are identified: productivity, multi-disciplinarity and author impact. As final step, the possibility to build a composite index, based on all dimensinos and all databases, will be critically evaluated.

References

Falagas M.E., Pitsouni E. I., Malietzis G. A. and Pappas G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strenghts and weaknesses. *The FASEB Journal*, 22, 338–342.

Marchant T. (2009). An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors *Scientometrics*, Vol. 80, No. 2 (2009) 327–344

[1] http://www.stat.unibo.it/ScienzeStatistiche/Ricerca/Progetti+e+attivita/Materiali_Giornata_di_Studio_-_La_valutazione_della_ricerca_nelle_scienze_statistiche.htm

Measuring the effect of cultural capital on students' university achievement

Isabella Sulis and Mariano Porcu

1 Introduction

This article moves from an *ad hoc* CATI survey carried out on students enrolled at the University of Cagliari in 2006/2007 academic year. The survey was addressed to detect the relationships between students' Cultural Capital (*CC*) (Bourdieu , 1994) at the end of the secondary school and students' universities choices and subsequent academic achievements. A special section of the survey questionnaire has been devoted to the measurement of the latent variable *CC* which has been operationalized including items addressed to know students and families habits.

Two main tasks have been pursued in this work: i) to operationalize the latent variable *CC* throughout the definition of its sub-components and ii) to assess the role played mainly by *CC* and others external factor (e.g., the support of institutions and family in the choice of the university pathway) in determining students' academic achievement. The main expected outcome is to isolate the role played by both factors in determining students' achievement.

2 Data and methods

This work is structured in two main parts. In the former the main task is to calibrate a set of key indicator items which allows to better shape the differences across students in terms of their *CC* and to determine clusters of students who applied different strategies of entrance at the university (Pitzalis et al. , 2008). In our research framework we assume by hypothesis that each individual possesses a basic amount of *CC*, namely the *inherited cultural capital* (CC_{IH}). It is evaluated by considering the highest level of formal education (measured in years of formal education in school or

Isabella Sulis, Mariano Porcu
Dip. Ric. Economiche e Soc. - Università di Cagliari, e-mail: {isulis,mrporcu}@unica.it

academic institutions) reached by students' parents. This basic amount of *CC* could be increased by family activities and habits (the *family made* CC_{FM}) and by the students' actions (the *pro-active* CC_{PA}). Moreover, the schooling process plays a role in the accumulation of *CC* and this factor is captured by the *formal education* (CC_{FE}) sub-component. The CC_{FM} and CC_{PA} are measured throughout activities made by students or by their families; the CC_{FE} is measured by 5 indicator variables (questionnaire items) addressed to show how the student evaluates her/his competencies in several subjects (math, language, literature, computer, overall competencies) and 2 objective indicators, one which marks the regularity of school career, the other signals if student attended or not a *Liceo* (the *Liceo* is a secondary school devoted to prepare students for university studies). Eight items were initially selected to scale the CC_{FM}, ten for the CC_{PA} and seven and the CC_{FE} component. The 25 indicator items take values on different scales: 10 are binary, 1 is a three-categories ordered, 14 are four-categories ordered. Indicator items to scale each sub-component have been re-calibrated by using an Item Response Model approach (Samejima , 1969) in order to: (i) select those items which effectively discriminate across subjects with a different magnitude of the latent trait; (ii) detect contiguous response categories on the ordinal scale that do not really indicate a different magnitude of the latent trait; (iii) assess the *difficulty* of the items in respect of the magnitude of *CC* owned by the surveyed population; (iv) measure students' *CC* considering the joint information available from the questionnaire. The regularity of the academic curricula has been measured considering the number of credits acquired three years after the enrollment.

In the second part quantile regression analysis has been adopted in order to explain sources of variability at each quantile of the achievement moving from the differences observed in the cultural components and in the strategies adopted in the choice of the university pathway. In this framework the method adopted: (i) allows to observe as the rate of change in the conditional quantile varies across quantiles; (ii) does not assume a functional form for the dependent variable and (iii) is particulary robust in presence of highly skewed distributions and outliers (Koenker and Hallock , 2001)

References

Bordieu P. (1994). Raisons pratiques. Sur la theorie de l'action. Edition du Seuil, Paris.

Koenker R., Hallock K. F.(2001). Quantile regression. *The Journal of Economic Perspectives*, 15(4), 143–156.

Pitzalis M., Sulis I., Porcu M. (2008). Differences of Cultural Capital among students in transition to university Some first survey evidences. *Quaderni di Ricerca Crenos*, 5.

Samejima F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17.

Formative measurement model for academic reputation

Emma Zavarrone

1 Introduction

In recent years the construct of corporate reputation has attracted more attention among researchers and practitioners. The present paper contributes to the emerging discussion on academic reputation literature by focusing both on the nature of the construct and on the stakeholders. According to Brewer and Zhao (2010) the academic reputation is a representation of past success and psychological status shared by stakeholders or networks in which individual members may not be the direct consumers of the offering. Then we can suppose that academic reputation depends on a number of stakeholders which engages relationship and transactions with the academy. The stakeholders under investigation are: government, research councils, academic researchers (as employees or collaborators), undergraduate students, families, companies. Furthermore, it is well known that the psychological status can be studied with latent methodologies. In this context we selected a formative measures in order to analyze the academic reputation. Shortly the use of formative indicators involves the creation of an index rather than a scale (Bollen and Lennox, 2001).In this context, formative measures are "observed variables that are assumed to cause a latent variable" (Bollen and Lennox, 2001).

2 Main results

The main research questions is: could we quantify the academic reputation trough a composite index? The answer for this question is based on the construction of a composite index using formative measures. The construction of an academic reputation index follows the guidelines suggested by Diamantopoulos and Winklhofer

Emma Zavarrone,
IULM,e-mail: emma.zavarrone@iulm.it

(2001): a) specification of the content, b) indicator or measures specification, c) indicator collinearity and d) external validity of indicators. These guidelines may be deemed necessary but do not help to discriminate the true nature of the measures used. (formative or reflective ones). For this reason we suggest the use of the tetrad vanish test proposed by Bollen (1991) an alternative to CFA for testing model fit.

The empirical part deals with survey: sampling choice (two stage sampling), administer questionnaire of 40 questions (measured score on continuous rating from 0 to 10) on academic reputation, collecting data (n=1680 students). Following Diamontopoulus' guidelines, we specify the domain of academic reputation as Performance, Governance, Innovation on Didactics, Products, college Fame. These dimensions have been proposed by author in a previous study. We select a pool of representative measures (10) from questionnaire items. These are linked to the previous aspects and with a low VIF value. The external validity of the chosen measures has been estimated with a Mimic model (ten formative measures and two reflective measures) Estimation of the proposed model produces a good fit (GoF=0.98 and RMSEA=0.014) even if not all ten indicators should be included in the index. The tetrad vanish test will be conducted using the macro CTANEST (SAS).

References

Brewer A., Zhao J. (2010). The impact of a pathway college on reputation and brand awareness for its affiliated university in Sydney. *International Journal of Educational Management*, 24(1), 34-47.

Bollen K. A., Lennox R. (2001). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin*, 110, 305-314.

Diamantopoulos A., Winklhofer H. (2001). Index construction with formative indicators: an alternative to scale development. *J Mark. Res.*, 37,269-77.

Contributed Session 2

Applications in Economics

A regression clustering method for the prediction of the pro capita disposal income in municipalities

Paolo Chirico

1 Introduction

The aim of *regression clustering* (Zhang, 2003) is segmenting a number of units in some clusters in order to detect a good regression model in each cluster. Then regression clustering is suitable when, given some explicative variables (regressors), a single regression model doesn't fit well all the units, but different regression models might fit well partitions of the data (see also Sarstedt and Schwaiger (2006)). In this paper a regression clustering procedure is adapted to a particular regression to predict the pro capita disposal income (PCDI) in municipalities. The particularity of this regression consist in: it is a two-level regression (municipalities and provinces) and the parameters estimation is run at the provincial level under some assumptions.

2 Main Results

The *PCDI* of a municipality is assumed explainable by some municipal indices (regressors) in a regressive model, but a single model for every municipality in a country or region would be a bit efficient: the regression errors may be too large. It is more flexible to assume the existence of K regressive models explaining the municipal *PCDI* in K clusters of provinces. So:

$$y_{ijk} = \mathbf{x}_{ijk}\beta_k + \varepsilon_{ijk} \tag{1}$$

where y_{ijk} is the *PCDI* of the i^{th} municipality in the j^{th} province of the k^{th} cluster; $\mathbf{x} = [1, X_1, X_2, ...]$ is the vector of the regressors and β is the vector of the correspondent parameters; ε is a random error. The distributional features of ε are inferred by the following model, assumed for the individual disposable income:

Paolo Chirico,
Dep. of Applied Statistics e Mathematics, Turin University, e-mail: paolo.chirico@unito.it

$$y_{hijk} = \mathbf{x}_{ijk}\beta_k + \varepsilon_{hijk} \tag{2}$$

where h indicates the individual; ε_{hijk} is a random error and includes all the individual factors determining y. Therefore it is assumed independent of each other error as well as of the regressors. No distributional form is assumed about ε_{hijk}, but only $E(\varepsilon_{hijk}) = 0$ and $Var(\varepsilon_{hijk}) = \sigma_k^2$.

As $y_{ijk} = \sum y_{hijk}/n_{ijk}$ then $\varepsilon_{ijk} = \sum \varepsilon_{hijk}/n_{ijk}$; generally $n_{ijk} > 1000$ so ε_{ijk} is approximately $N(0, \sigma_k^2/n_{ijk})$. Moreover ε_{ijk} is independent of each other error and of the regressors as well. Consequently the model for the provincial PCDI is:

$$y_{jk} = \mathbf{x}_{jk}\beta_k + \varepsilon_{jk} \tag{3}$$

where $y_{jk} = \sum y_{hijk}/n_{jk}$ and $\varepsilon_{jk} = (\sum \varepsilon_{hijk}/n_{jk}) \sim N(0, \sigma_k^2/n_{jk})$.

The *PCDIs* of the municipalities are unknown (they are the object of the prediction), but the *PCDIs* of the provinces are. So the parameters in β_k are estimated through provincial data by the *WLS* method.

The clusters are determined by a segmentation oriented to the efficiency of the local regression models. This procedure of "regression clustering" (see Introduction), is a model-based version of the K-means clustering method. It is characterized by the following steps:

1. estimation of a global regression model on all provinces;
2. hierarchical classification on the residual of the global model (dendogramme);
3. choice of the number, K, of the clusters according to the dendogramme and assignment of provinces to the K clusters;
4. estimation of the K local model (one for each cluster) and computation of the prediction error for each provinces in each K local model;
5. assignment of each provinces to the closest local model (where the prediction error is the smallest);
6. repetition of the steps 4. and 5. until the composition of the K clusters doesn't change.

References

Sarstedt M., Schwaiger M. (2006). Model Selection in Mixture Regression Analysis: A Monte Carlo Simulation Study. *Data Analysis Machine Learning and Applications*, Springer Berlin Heidelberg, 61-68.

Zhang B. (2003). Regression Clustering. In: *ICDM03, Third IEEE International Conference on Data Mining*, 451.

Discrete and continuous time mover-stayer model for labour market in a small northern Italian area

Fabrizio Cipollini, Camilla Ferretti, Pero Ganugi, Mario Mezzanzanica

1 Introduction

Through the law 196/97 (”pacchetto Treu”) and other legislative measures Italian labor market acquires a degree of flexibility that is unknown in the previous two decades. Working relations are in fact codified in 35 contracts characterized by different degree of guarantees. Using the COB archive of Cremona we aim to investigate the mobility among contracts, in particular from those assuring a very modest package of rights toward more structured working relations ending with full-time (unlimited duration) contracts.

Data include all people working in Provincia of Cremona in both January 2007 and September 2009. The statistical tool we use is the *Mover-Stayer Model* (MS). Applications of the model, among others, are in Cipollini et al. (2009) (discrete time version) and in Fougere and Kamionka (2003) (continuous time version). We group the 35 contracts in eight possible states, ordered according to package of guarantees (table 1).

Table 1 The states.

1. Unlimited duration and Full time	} Dependent Contracts
2. Unlimited duration and Partial time	
3. Expiry job and Full time	
4. Expiry job and Partial time	
5. Apprenticeship	
6. Co.co.pro. and Co.co.	} Parasubordinate Contracts
7. Self-employment	} ”Autonomo”
8. Unemployment	} No Contracts

Fabrizio Cipollini,
Dept. of Statistics, Università di Firenze, ITALY, e-mail: cipollini@ds.unifi.it

Camilla Ferretti, Pero Ganugi,
Dept. of Economics and Social Sciences, Università Cattolica del Sacro Cuore, Piacenza, ITALY,
e-mail: camilla.ferretti@unicatt.it, piero.ganugi@unicatt.it

Mario Mezzanzanica,
Dept. of Statistics, Milano Bicocca, ITALY, e-mail: mario.mezzanzanica@unimib.it

MS model subdivides units in two groups: every individual starting from state i can be a *stayer* (with probability s_i) or a *mover* (with probability $1-s_i$). In the first case he never moves; in the second one he moves among states according to a Markov Chain (MC) with transition matrix M. We denote $\theta = (s, M)$ and we assume that the MC is a continuous time process. The corresponding generating matrix Q is estimated by means of the *Gibbs Sampler*.

2 Main results

Our procedure can be summarized:

1. from data we derive the starting distribution n_0, the number N_{ij} of observed transitions from i to j for every i and j, and the number n_{si} of observed workers continuously remaining in i;
2. introducing the unobservable vector Z, such that Z_i is the number of stayers in i, we show:

$$Z_i|\theta \sim \text{Binomial}(n_{si}, p_i) \text{ with } p_i = p_i(s_i, M_{ii}),\ \forall i = 1, \ldots 7 \tag{1}$$

$$\theta|Z \sim \text{MatrixBeta}(\{A_{ij}\}_{i,j}) \text{ with } A_{ij} = A_{ij}(n_0, N_{ij}, n_s, Z) \tag{2}$$

3. using (1) and (2) we implement $n = 10000$ iterations of Gibbs sampler aiming to demonstrate the existence of the continuous-time process and to estimate Q;
4. we use the estimated $\bar{Q}$ to obtain the equilibrium distribution.

Comparing starting and equilibrium distributions we note a relevant decrease of mass in unlimited-full-time state and apprenticeship. At the opposite we note an increase of the mass in the unlimited-partial-time and expiring-partial time states.

Table 2 Starting versus Equilibrium distribution (calculated with $\bar{Q}$)

State	1	2	3	4	5	6	7	8
d_0	54.94%	18.90%	9.60%	7.69%	6.57%	1.15%	0.07%	1.08%
d_{eq}	48.64%	25.08%	10.82%	9.20%	2.59%	1.01%	0.13%	2.52%

References

Cipollini F., Ferretti C., Ganugi, P. (2009). Mover-Stayer Model in an Indus- trial Area: a rst application. In: *Statistical Methods for the Analysis of Large Data Sets*, 355–358.

Fougere, D. and Kamionka, T. (2003). Bayesian Inference for the Mover-Stayer Model of Continuous Time, *J. Appl. Econ.*, 18, 697723.

Electricity consumption and gross domestic product in the Italian regions

Antonio Angelo Romano and Giuseppe Scandurra

1 Introduction

This paper examines the relationship between economic growth of the Italian regions and electricity consumption. In particular, we investigate the nature of the link between changes in electricity consumption in Italy and the variation of its GDP. It is well knowns, in fact, that there are different assumptions about the nature of economic growth in relation to energy consumption in general and the electricity consumption in particular. Electricity consumption is an important percentage of total energy consumption thanks to the network infrastructure significantly extended in a country like Italy and, of course, the fact that its measure is characterized by extreme accuracy, timeliness and spatial granularity. Given the economic heterogeneity of the Italian regions, it would be reasonable to expect the relationship between energy consumption and economic growth that in some cases support the hypothesis of neutrality (Yu and Choi, 1985) that the energy consumption should not significantly affect the economic growth of one region and other cases in which energy consumption plays, however, a major role. From the statistical point of view is whether there is a unidirectional causality of energy consumption on income, or if that link takes on the characteristic bidirectional as a sign of complementarity between the two variables.

Antonio Angelo Romano
Department of Statistics and Mathematics for Economic Research, University of Naples "Parthenope", 40, Medina - Naples - I - 80133, e-mail: antonio.romano@uniparthenope.it

Giuseppe Scandurra
Department of Statistics and Mathematics for Economic Research, University of Naples "Parthenope", 40, Medina - Naples - I - 80133, e-mail: giuseppe.scandurra@uniparthenope.it

2 Main results

This paper applies recent advances in panel analysis to estimate the panel cointegration and panel vector error correction models for the Italian regions using annual data covering the period 1980–2008. Our cointegration and causality analyses can be summarized in the following steps:

1. test of heterogeneity among the regional units: according to the test of Breush and Pagan based on Lagrange multipliers, it was possible to verify the heterogeneity of the regional units.
2. the existence of parameter heterogeneity suggests the use of Im et al. (2003) panel unit root test to determine the integration order of the series. From this point of view, it was verified that the variables have both a stochastic trend. Their first difference is stationary. So, GDP and electricity consumption are integrated of order one.
3. cointegration test between the two series: for this purpose was used the procedure proposed by Westerlund (2007) because it is considered more robust than the Pedroni's cointegration test in the case of small samples. The test confirms the presence of a long–run equilibrium relationship between the series analyzed.
4. given the presence of cointegration, the Fully Modified Ordinary Least Squares (FMOLS) technique (Pedroni , 2000) is estimated to determine the long–run equilibrium relationship.
5. to infer the causal relationship between the variables we estimate a panel vector error correction model based on the generalized method of moments (GMM) estimator proposed by Arellano and Bond (1991).
6. finally, estimates FMOLS effected for each region, allowed a first evaluation of the different characteristics of each in relation to the sign and size of the coefficients of the model.

References

Arellano M., Bond S.R., (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277-297.

Im K. S., Pesaran M. H., Shin Y., (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115, 53-74.

Pedroni P. (2000). Fully Modified OLS for the heterogeneous Cointegrated Panels. *Advances in Econometrics*, 15, 93-130.

Westerlund J. (2007). Testing for Error Correction in Panel Data. *Oxford Bulletin of Economics and Statistics*, 69 (6), 709-748.

Yu E. S. H., Choi J. Y. (1985). The causal relationship between energy and GNP: an international comparison. *Journal of Energy and Development*, 10, 249-272.

Contributed Session 3

Dimension Reduction

Variable selection in a predictive approach

Simone Borra, Agostino Di Ciaccio

1 Introduction

The notion of what makes a variable "important" in a model is controversial, but one interpretation is that a variable is important if dropping it seriously affects prediction accuracy (Breiman, 2001). In this paper we compare some methods for variable selection in a predictive approach, considering also the interpretation capability of the models. In fact, for real sample size, there is a significant gap between the "predictive approach" and the approach looking for the "true model". The result is that an optimal model for prediction purposes may differ from the one obtained looking for the "true model". As noted by Akaike (1985) "Except for data obtained by an artificial sampling scheme, we do not know exactly what is meant by the true distribution. Indeed, the concept of the true distributions obtains a practical meaning only through the specification of an estimation procedure or a model."

2 Main results

In this paper, to make a comparison between variable selection methods, we assume that the "true" data generation process is $Y = f(X_1, X_2 ....., X_J) + \varepsilon$, where f is a unknown function of J covariates $X = X_1, X_2 ....., X_J$ and ε is a random noise with mean zero and variance σ^2. We can estimate the function $f(X)$ using a training-set $\mathbf{c}$, with K covariates ($K > J$) obtaining $\hat{f}_c$. To select the most 'important' variables in a predictive approach, we should know the prediction capability of $\hat{f}_c$ for each combination of variables. Of course, the relevance of each variable will depend

Simone Borra,
Dept. DET, Univ. of Rome "Tor Vergata", e-mail: borra@economia.uniroma2.it

Agostino Di Ciaccio,
Dept. of Statistics, Sapienza Univ. of Rome e-mail: agostino.diciaccio@uniroma1.it

on the predictor model and on the sample drawn. Chosen a loss function L(.), the Prediction Error of a fixed $\hat{f}_c$, is given by

$$Err = E_{x_0} E_{Y_0}[L(Y_0, \hat{f}_c(x_0)) \mid \hat{f}_c] \quad (1)$$

which is named Extra-sample error and measures the expected loss of $\hat{f}_c$ on the population (note that $\hat{f}_c$ depends on the training sample). A more restrictive definition of Prediction Error is the In-sample error, where the values of the covariates are considered fixed at their observed sample values X_c while Y is random.

For variable selection, we can use the classical information criteria AIC, the Bayesian BIC, an estimator of Err_{in}(for example using Parametric Bootstrap) or, finally, estimate directly the Extra-sample error. In this case K-fold Cross-Validation or Bootstrap 632+ could be appropriate. A recent penalized approach in linear regression problems is given by the LASSO and LAR models (Efron et al., 2004).

Through the use of an extensive simulation study we compared the performance of these different approaches. In particular we considered multinormal and non-multinormal data with multiple linear regression and regression trees. We generated samples considering 10 quantitative covariates while Y is a function of J$<$10 variables, plus a normal noise. The generating function is linear in the first simulation and non-linear in the second simulation, with several levels of signal to noise ratio. We generated many populations, and for each population we drew 300 samples. The analysis was repeated for several sample-sizes. The "true" value of Err was obtained generating another large test-set of 50,000 units.

To evaluate the different methods we calculated the *accuracy index*, which measures the capability of the method to individuate the "true" model", and the *mean relative error*, which compares the Prediction Error of the estimated model with the P.E. of the estimated model including only the "true" variables. Analysing the obtained results, it is evident that using accuracy or Prediction Error we can obtain different models. LASSO detected models with a good performance in the multinormal case. In the non-linear case, K-fold CV and Parametric Bootstrap can individuate models which give better predictive capability than models obtained considering the 'true' variables.

References

Akaike H (1985). Prediction and entropy. *A Celebration of Statistics*, A.C. Atkinson & E. Fienberg eds., Springer-Verlag, New York, 1-24.

Breiman L. (2001). Statistical modeling: the two cultures. *Statistical Science*, 16: 199-215.

Efron B., Hastie T., Johnstone I., Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, 24:407-499.

Adaptive linear dimension reduction in a classification setting

Karsten Luebke, Claus Weihs

1 Introduction

The no-free-lunch theorem (Wolpert , 2001) states that there can not be a best classification method for all possible data. So the question is: Which method is best for the data at hand. The StatLog (Michie, Spiegelhalter, Taylor , 1994) and METAL (Brazdil, Soares, Pinto da Costa , 2003) project for example are investigating a number of data sets and try to generate rules based on attributes or meta features of the data sets (e.g. number of observations). This metalearning (Brazdil, Giraud-Carrier, Soares, Vilalta , 2009) can be important for practitioners as under time pressure (e.g. classification for direct marketing) when there is no time for evaluating the performance of let's say 10 different classifiers by means of cross-validation. Quite a lot of the proposed classification and dimension reduction methods in the literature are evaluated by some benchmark data sets (with Fisher's Iris Data being the oldest) – preferably by those data sets in which the new method performs well. Unfortunately the data situations (attributes or assumptions) in which it is beneficial to use a proposed method are left unclear by this.

2 Main results

Our adaptive procedure is a further development of the aforementioned metalearning as it constructs rules based on simulated data which are aimed to cover the

Karsten Luebke,
Hochschule für Oekonimie und Management, c/o B1st software factory, Rheinlanddamm 201, 44139 Dortmund, Germany, e-mail: karsten.luebke@fom.de

Claus Weihs,
Dortmund University of Technology, Department of Statistics, 44221 Dortmund, Germany e-mail: weihs@statistik.uni-dortmund.de

space of possible data situations. The actual method used to classify the data at hand depends on this data and is selected by means of a so-called selection statistic (O'Gorman , 2004). The task of finding the statistic is mainly the task of an adequate description of the data. The adaptive procedure can be described by the following steps:

S1 Identification of data space determined by feasible data attributes.
S2 Calculation of selection statistic (decision rule) for all possible data situations.
S3 Application of selection statistic to the given data situation.
S4 Application of selected method to the data at hand.

The aim of our work is twofold: first we want to introduce the adaptive procedure for classification settings, on the other hand we are considering data attributes influencing the performance of methods for linear dimension reduction in classification.

In the space of data situations local three-class scenarios can be found in which the performance gain is substantial and an adaptive procedure can be found using a simple rule – for all possible data situations under the given assumptions.

For more realistic scenarios the selection statistic depends on many more data characteristics. Nevertheless rules can be found for an adaptive procedure based on a pre-analysis of the data that can improve classification for real life data.

Therefore, it may be worthwhile for practitioners to check the possible data situations for their every-day work and do this meta-learning rather than a time-consuming trial-and-error approach.

References

Brazdil, P., Soares, C. and Pinto da Costa, J. (2003). Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results. *Machine Learning*, 50 (3), 251–277.

Brazdil, P., Giraud-Carrier, C., Soares, C. and Vilalta, R. (2009). *Metalearning*, Berlin, Heidelberg: Springer.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994). *Machine learning, neural and statistical classification*, New York: Ellis Horwood

O'Gorman, T. W (2004). *Applied Adaptive Statistical Methods: Tests of Significance and Confidence Intervals*, Philadelphia: SIAM

Wolpert, D.H. (2001). *The Supervised No-Free-Lunch Theorems*, In: Proceedings of the 6th On-line World Conference on Soft Computing in Industrial Applications, 25–42

Dimensional reduction and clustering of functional data

Roberto Rocci, Stefano Antonio Gattone

1 Introduction

When data are functions, reduction methods, like cluster or principal component analysis (PCA), have to cope with the infinite dimensional setting of functional data. In clustering, a common way to proceed is to filter first by means of a finite number of basis functions and then cluster the basis coefficients (Abraham *et al* (2003)). Some other authors propose to work on the original discretized data by imposing some form of regularization to take into account the functional nature of the data (Chiou and Li, 2007). In the filtering approach, smoothing is implied by the truncated basis chosen (presmoothing) while in the regularization one it is contextual to the analysis (contextual smoothing). Furthermore, functional PCA (Ramsay and Silvermann (2005)) could be applied to improve the clustering results by reducing the data dimension (Song *et al*, 2007). This two-step technique is often used by practitioners but criticized by several authors, because the PCA reduction step could remove some significant information about the clustering structure of the data. The aim of this work is to propose a new method, combining PCA and clustering, where the classification is obtained in a lower dimensional subspace. The method is implemented by following both the pre and contextual smoothing approaches.

2 The Method

Let $\{x_i(t)\}_{i=1}^{I}$ be a set of functional observations where $t \in \Gamma \subset \Re$. The aim is to classify the sample $\{x_i(t)\}_{i=1}^{I}$ into G homogeneous groups. The data are represented

Roberto Rocci,
SEFeMeQ, University of Tor Vergata, Rome e-mail: roberto.rocci@uniroma2.it

Stefano Antonio Gattone,
SEFeMeQ, University of Tor Vergata, Rome e-mail: gattone@economia.uniroma2.it

in the following signal plus noise setup

$$x_i(t) = \sum_{i,g} u_{ig} m_g(t) + \varepsilon_i(t) \qquad i = 1,...,I; g = 1,...,G; \tag{1}$$

where m_g are smooth unknown centroid functions, $\varepsilon_i(t)$ denotes an unobservable zero-mean error term and $u_{ig} = \{0,1\}$, $\sum_g u_{ig} = 1 \ \forall i$ and $u_{ig} = 1$ if x_i belongs to the g-th cluster. The estimation of the u_{ig} and m_g can be carried out by minimizing the within clusters deviance

$$\sum_{i,g} u_{ig} \int_T [x_i(t) - m_g(t)]^2 dt = \sum_{i,g} u_{ig} \left\| x_i - m_g \right\|^2 \to min. \tag{2}$$

Model (1) is modified by assuming that the centroids m_g lie into a Q dimensional subspace spanned by the smooth functions $\{\psi_q(t)\}_{q=1}^{Q}$, i.e. m_g may be expressed as

$$m_g(t) = \sum_q a_{gq} \psi_q(t). \tag{3}$$

Estimates of a_{gq} and $\psi_q(t)$ can be obtained by minimizing (2) under the constraint (3) so to find a low dimensional representation of the data which provide the largest separation among the clusters. This approach does not make any use of the smoothness of the m_g's. To incorporate the smoothing step in the cluster analysis, we propose to add a roughness penalty to the criterion (2)

$$\sum_{i,g} u_{ig} \left\| x_i - m_g \right\|^2 + \lambda \sum_g \left\| m_g'' \right\|^2 \to min \tag{4}$$

where m_g'' is the second derivative of m_g. A key factor is the proper selection of the smoothing parameter λ that controls the trade-off between the variance and the roughness of $m_g(t)$. We suitably adapt the generalized cross-validation (GCV) criterion in order to properly select the amount of smoothing.

References

Abraham C., Cornillon P. A., Matzner-Lber E., Molinari N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, 30, 581-595.

Chiou J. M., Li P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society, Series B*, 69, 679-699.

Ramsay J. O., Silverman B. W. (2005). *Functional Data Analysis* (2nd ed.), New York: Springer-Verlag.

Song J. J., Leeb H. J., Morris J. S., Kang S. (2007). Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*, 31, 265-274.

Contributed Session 4

Environmental Data Analysis and Classification

Functional clustering of temperature and precipitation data for Italian climate zones determination

Edmondo Di Giuseppe, Giovanna Jona Lasinio, Massimiliano Pasqui, Stanislao Esposito

1 Introduction

In order to correctly describe atmospheric variability and clear trends, homogeneous climate regions should be identified. A set of different methods are used for climate zones determination, typically a combination of Principal Component Analysis (PCA) and Cluster Analysis (CA).

Using time series of meteorological observations at various locations PCA is applied in the spatial domain (S-mode PCA) the aim being a regionalization of the study area (Ehrendorfer, 1987). On the other hand a temporal mode (T-mode) approach can be developed with a CA in order to look at locales similarity in mean and variances of temperature and precipitation across a fixed time (Fovell and Fovell, 1993).

In this work a combination of Functional Data Analysis (FDA) and Partioning Around Medoids (PAM) clustering technique is applied in Italy for surface temperature and precipitation fields in order to delineate climate zones for Agroscenari research project. The main advantage of a functional approach to this type of data is dimensional reduction, as the information on temporal pattern given by a large number of observations (time series) is summarized by few coefficients describing the basis spanning the chosen functions (Ramsay and Silverman, 1997). B-splines system of basis with a fixed number of knots is adopted for converting observations gathered at discrete time into functional data. Finally the Pcs of estimated coeffi-

Edmondo Di Giuseppe,
CRA-CMA and Univ. of Rome La Sapienza, e-mail: edmondo.digiuseppe@entecra.it

Giovanna Jona Lasinio,
University of Rome La Sapienza-Dspsa,e-mail: giovanna.jonalasinio@uniroma1.it

Massimiliano Pasqui,
CNR-Ibimet,e-mail: m.pasqui@ibimet.cnr.it

Stanislao Esposito,
CRA-CMA,e-mail: stanislao.esposito@entecra.it

cients are partitioned by PAM classification technique and average silhouette width method is used to determine the number of climate zones (Rousseeuw, 1987).

2 Main results

The dataset is composed of daily precipitation and daily minimum and maximum temperature data collected for the period 1961-2007 from 96 Italian stations. First, minimum and maximum temperatures were averaged to obtain daily medium temperatures. Then Monthly Mean of Medium Temperature (Tmed-MM) and Monthly Cumulated Rainfall (Prec-MC) were calculated. Thus, 96 time series of 564 monthly values concerning a set of 2 climatic variables form the basis for the classification.

For Tmed-MM a graphical analysis suggests to use a knot every six months, with 94 fixed knots and 3 piece-wise polynomials degree. Varimax rotated Pc1 and Pc2 of 96 estimated coefficients accounts for 98.5 % of total variability and an average silhouette width value of 0.58 leads to decide for a 3 clusters partition.

For Prec-MC the best model is quarterly with 188 fixed knots and 3 piece-wise polynomials degree. A PCA on 190 estimated coefficients accounts for 69.6% of total variance on first 4 Pcs and after a Varimax rotation, the silhouette width plot reveals a not well defined partition structure. This suggests the requirement of further development.

Table 1 Tmed-MM and Prec-MC selected models for functional data transformation with B-splines order, number of knots, penalty coefficient (lambda) and RMSE.

Variables	Selected Model	Order	Knots	Lambda	Rmse
Tmed-MM	six-monthly	4	94	0.06	1.916
Prec-MC	quarterly	4	188	2.51	0.519

References

Ehrendorfer M. (1987). A regionalization of Austria's precipitation climate using principal component analysis. *International Journal of Climatology*, 7, 1, 71-89.

Fovell R.G. and Fovell M.Y.C. (1993). Climate Zones of the Conterminous United States Defined Using Cluster Analysis. *J. Climate*, 6, 2103-2135.

Ramsay J. O., Silverman B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

Rousseeuw P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

The application of M-function analysis to the geographical distribution of earthquake sequence

Eugenia Nissi, AnnaLina Sarra, Sergio Palermi, Gaetano De Luca

1 Introduction

Many statistical studies show that earthquake distribution in a region is usually dominated by significant clusters in both space and time. The lack of spatial independence in seismic data is traditionally perceived as a problem obscuring the ability to separate the background seismicity from clustering pattern. Accordingly a relevant role in the study and the comprehension of seismic process is played by the second-order property of the point processes. The well-known Ripley K-function describes the degree to which there is spatial dependence in the arrangement of events and is a powerful tool to assess whether or not a point pattern satisfies the Poisson distribution. In this paper we explore a variant of Ripley's K function (Ripley, 1976), the M function, as a new means of quantifying the clustering of earthquakes. In particular we test how the positions of epicentres are clustered in space with respect to their attributes values, i.e the magnitude of the earthquakes. The strength of interaction between events is discussed and results for L'Aquila earthquake sequence are analysed.

2 Main results

The M function constitutes a generalisation of Ripley's K function, developed by Marcon and Puech (2003). Through the M function, we are able to compare the number of a certain type of event (say N_{mk}) to the total number of event (say N). In our context, the Marcon and Puech solution implies the definition of the following

Eugenia Nissi,
Department of quantitative Methods and Economic Theory, e-mail: nissi@unich.it

AnnaLina Sarra,
Department of quantitative Methods and Economic Theory, e-mail: a.sarra@dmqte.unich.it

steps: a circle with radius r is drawn around each epicentres with a predefined magnitude, then not only is the number of epicentres with a certain magnitude counted in each of those circles, but also the number of all epicentres. Those two numbers are then used to calculate the quotient for each event (epicentres with a certain magnitude); those quotients are then used to calculate the average over all those circles. Finally that average is divided by the total number of events in the study area as a whole. So if the focus is on the research of a particular neighbour, the Ripley's K function can be easily adapted for that purpose and the formula to be used is the following:

$$M'_{mk}(r) = \frac{\frac{\sum_{J=1}^{N_{mk}} \frac{\sum_{i=1,i\neq j}^{N_{mk}} I_{mk}(i,j,r)}{\sum_{i=1,i\neq j}^{N} I_{mk}(i,j,r)}}{N_{mk}}}{\frac{N_{mk}-1}{N-1}} \quad (1)$$

where $I_{mk}(i,j,r)$ stands for a dummy: its value is 1 if the point j is located inside the circle and 0 otherwise (more details can be found in Marcon and Puech (2003)). The earthquake sequence investigated belongs to the area of Central Italy (L'Aquila) bounded in longitude by 13.034 and 13.749 degrees East and in Latitude by 42.127 and 42.634 North. The events pertain to a time period spanning from 1st October 2008 to 30 th October 2009. The earthquake process is characterized by spatial clustering as revealed by Ripley's K function which exhibits spatial aggregation under 25 Km. As we are interested in assessing the foreshock and aftershock activities, in particular in their spatial interaction, the focus here is on quantifying the scale at which clustering of epicentres with magnitude greater than 2.5 and 3 takes place, by means of M function. The mainshock of magnitude 5.9 occurred on 6 April 2009 splits our database in two parts: the foreshock and aftershock sequences. The M function analysis for the foreshocks sequence, where events with magnitude greater than 3 are investigated, reveals significant concentration of this kind of events up to 8 Km. By contrast the M values of the aftershock sequence show important concentration peaks at larger distances, i.e in a radius of about 30 Km. For the weaker foreshock sequence (events with magnitude greater than 2.5) the spatial concentration measured by the M function is observed at a radius less than 500 meters. Always on the basis of M function analysis we again observe a valuable concentration under 30 km for the aftershock sequence with magnitude greater than 2.5.

References

Ripley B.D. (1976). Second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2), 255-266.

Marcon E., Puech F. (2003). Measures of the Geographic Concentration of Industries: Improving Distance-based Methods *Working paper. Université Paris I, Cahiers de la MSE*.

From a multivariate spatio-temporal array to a Multipollutant - Multisite Air Quality Index

Antonella Plaia, Mariantonietta Ruggieri, Francesca Di Salvo, Gianna Agró

1 Introduction

In recent years, quantifying air quality and, most of all, following the evolution of pollution, has been becoming a fundamental issue for local and central governments. The need to inform the population about the air quality, and related health outcome, has led to a proliferation of synthetic indices (Air Quality Indices - AQIs), giving an idea of the state of pollution in a day (Plaia and Ruggieri, 2002). On the other side, an AQI time series, which syntetizes air quality with respect to different pollutants and many monitoring sites, can be used to follow the evolution of air quality in a town/region.

AQIs are computed on air pollution data that are usually collected according to time, space and type of pollutant: in a given town/region, data consisting of hourly levels of K pollutants recorded in S monitoring sites, are usually organized in a three-mode array. A first aggregation step usually concerns time, and allows to pass from hourly data to a daily synthesis: in this paper data will be aggregated by time according to the guidelines provided by the national agencies (Ruggieri et al., 2009) producing the three mode array $\mathbf{X}_{[T\times S\times K]}$. Here we will propose a new approach to get a Multipollutant-Multisite Air Quality Index time series from a multivariate spatio-temporal array. This implies a two step aggregation, according to space and to pollutant. Since the value of the synthetic index depends on the choice of the aggregating functions and the order of aggregation, here we will follow a monitoring site - pollutants aggregation (Bruno and Cocchi, 2002), and:

- in order to find a spatial synthesis of pollutant time series a three mode factor analysis will be used, obtaining AQI_k;
- the AQI (I_2), proposed by Ruggieri et al. (2009) with respect to a single monitoring site, will be then applied to represent the global air pollution situation on a

Antonella Plaia, Mariantonietta Ruggieri, Francesca Di Salvo, Gianna Agró
Department of Statistical and Mathematical Sciences, University of Palermo, viale delle Scienze - building 13, 90128 Palermo, Italy,
emailplaia@unipa.it, ruggieri@dssm.unipa.it, disalvo@dssm.unipa.it, agro@unipa.it

given day:

$$I_2 = (\sum_{k=1}^{K} (AQI_k)^2)^{\frac{1}{2}}. \quad (1)$$

2 Main results

Several decomposition methods of multi-way data can be found in literature, that treat the three modes symmetrically, like PARAFAC or TUCKER3 (or three-mode Principal Component Analysis) or asymmetrically like STATIS or unfolding PCA (Kiers, 2000).

Here an asymmetric view is adopted reordering the three-way data in order to focus on space rather than on time as usually done. The three way array with elements x_{tsk} with $t = 1,2,\ldots T$, $s = 1,2,\ldots S, k = 1,2,\ldots K$, is unfolded (matricized) in a matrix $\mathbf{Y}_{[TK\times S]} = [vec(\mathbf{X}_1), vec(\mathbf{X}_2), \ldots, vec(\mathbf{X}_S)]$ where $\mathbf{X}_s$ is the T by K matrix corresponding to data recorded at monitoring site s, while $vec(\mathbf{X}_s)$ denotes the TK vector containing all the elements of $\mathbf{X}_s$ strung out rowwise into a column vector.

In order to compare different pollutants having different measurement units or order of magnitude, a standardization has to be performed (preprocessing of data): a standardizing transformation by linear interpolation, that takes into account also effects on human health occurring over long time periods should be preferred since, in such a way, classes for low concentrations of a pollutant are also considered (see (Ruggieri et al., 2009) for details about standardizing function and class breakpoints). After preprocessing of data, *Tucker 1 model* (Kiers, 2000) is applied in order to get a reduction in only one mode (space). Finally I_2 is applied in order to represent the global air pollution situation in a city on a given day. The method will be applied to air quality data recorded in Palermo (Italy). The three way array contains the levels of five pollutants (CO, NO2, O3, PM10, SO2) measured hourly at nine monitoring sites over a 1-year period (2006). I_2 behaviour along the year will be compared to other methods in literature, in particular to the set of AQIs proposed in (Bruno and Cocchi, 2002).

References

Bruno F., Cocchi D.(2002). A unified strategy for building simple air quality indices. *Environmetrics* 13, 243-261.

Kiers H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of chemometrics* 14, 105-122.

Plaia A., Ruggieri M. (submitted). Air quality indices: a review. Submitted.

Ruggieri M., Plaia A., Bondí A.L. (2009). Aggregate air pollution indices: a new proposal. In: Classification and Data Analysis 2009, Book of Short Papers, Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, 221-224.

Contributed Session 5

Applications in Demography

Analysis of the individual variability of sex ratio with hierarchical models

Silvia Bozza, Mario Di Bacco, Renzo Bigazzi, Sergio De Iasio

1 Introduction

The propensity of human couples to generate children of the same sex has been object of several studies. An early answer has been provided by C. Gini in his book (Gini, 1908), where a moderate individual variability of sex ratio was supported. Subsequent studies (e.g., James (1987)) have focused the attention on the influence that several variables such as ethnic group, smoking, blood group etc., might have to the value of sex ratio, but none of them could provide an effective support in favor or in contrast of this hypothesis. A moderate differential propensity was detected by Rinaldi et al. (2003). Starting from a database of families reconstructed from civil and parochial certificates of birth, marriage and deaths, the authors implemented a Bayesian hierarchical model (Gelman et al., 2003) to infer the magnitude of individual variability. This paper develops further this model in the sense that two generations are considered. Numerical procedures are implemented to handle the complexity of the posterior distributions.

Silvia Bozza,
Dipartimento di Statistica, Università Ca' Foscari, Venezia, e-mail: silvia.bozza@unive.it

Mario Di Bacco,
Scuola di Alta Formazione Statistica, Polo Universitario di Asti, e-mail: biostat@uni-astiss.it

Renzo Bigazzi,
Facoltà di Medicina e Chirurgia, Università di Firenze, e-mail: renzo.bigazzi@unifi.it

Sergio De Iasio,
Dipartimento di Genetica, Biologia dei Microrganismi, Antropologia, Evoluzione, Università di Parma, e-mail: sergio.deiasio@unipr.it

2 Main results

The database is given by $N = 233$ families reconstructed using the biodemographic study of the population of Alia (Palermo) that was based on civil certificates of birth, marriage and death from 1815 to 1899. For each couple j in the database, there are available the family of the paternal and of the maternal grandfather. The paternal grandfather had n_{1j} children (of which, x_{1j} are males), while the maternal grandfather had n_{2j} children (of which, x_{2j} are males). Each couple had T_j children (of which, t_j are males). Then the likelihood $l(x \mid \theta_{1j}, \theta_{2j})$ for each couple is

$$l(x \mid \theta_{1j}, \theta_{2j}) \propto \left(\frac{\theta_{1j} + \theta_{2j}}{2}\right)^{t_j} \left(1 - \frac{\theta_{1j} + \theta_{2j}}{2}\right)^{T_j - t_j} \prod_{i=1}^{2} (\theta_{ij})^{x_{ij}} (1 - \theta_{ij})^{n_{ij} - x_{ij}}, \quad (1)$$

where θ_{ij} denotes the probability of 'male' for the paternal ($i = 1$) and the maternal ($i = 2$) grandfather, while the probability of 'male' for the second generation is taken to be the mean of the respective probabilities inherited from the paternal and maternal grandfather. Assume that probabilities θ_{ij} are not necessarily equal to $1/2$. To model the uncertainty about these values, a beta distribution $\pi(\theta_{ij})$ with parameters $\alpha = \beta = \tau$ is introduced, such that $E(\theta_{ij}) = 1/2$, while the variability is related to the value of the hyperparameter τ. The object of inference becomes the variance of the beta distributed variable, since the more pronounced will be this variability a posteriori, the more data will support the presence of differential propensity. A prior distribution $\pi(\tau)$ is therefore introduced. Given the N families, the likelihood is

$$L(x \mid \tau) = \prod_{j=1}^{N} \int_0^1 \int_0^1 l(x \mid \theta_{1j}, \theta_{2j}) \pi(\theta_{1j}) \pi(\theta_{2j}) d\theta_{1j} d\theta_{2j}. \quad (2)$$

The posterior distribution

$$\pi(\tau \mid x) \propto L(x \mid \tau) \pi(\tau) \quad (3)$$

is not tractable in closed form and will be obtained numerically. The degree of dispersion in the posterior distribution will be analyzed.

References

Gelman A., Carlin J. B., Stern H. S., Rubin D. B. (2003). *Bayesian data analysis*, 2nd edition. Chapman & Hall/CRC.

Gini C. (1908). *Il sesso dal punto di vista statistico*, Sandron, Milano.

James W. H. (1987). The Human sex ratio. Part 2: A hypothesis and a program of research. *Human Biology*, Waine State University Press.

Rinaldi A., Bigazzi R., De Iasio S., Di Bacco M. (2003). Individual variability of sex ratio: a statistical control. *Genus*, 59, 29-36.

Clustering of population pyramids presented as histogram symbolic data

Simona Korenjak-Černe, Nataša Kejžar, Vladimir Batagelj

1 Introduction

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. Its shape is influenced not only by fertility, mortality and migrations as the main demographical indicators, but also by many other social and political policies and events, such as birth control policy, wars, life-style etc. As such, the shape of the pyramid shows many demographic, social and political characteristics of the time and the region. Therefore, clusters of countries with similar pyramidal shapes can offer additional insight into the study of countries for demographers, sociologists, and other researchers.

With the adapted hierarchical clustering procedure we want to inspect the clusters of similar countries based on the shapes of their population pyramids. The changes of the pyramids' shapes, and also changes of the countries inside main clusters will be examined for the years 1996, 2001, and 2006.

2 Main results

Data on the population pyramids of the world countries used in our analysis were taken from the web page of the International Data Base (IDB). Age is divided into 17 five-years groups (0-4 years, 5-9 years, 10-14 years, ..., 75-79 years, 80+). In a

Simona Korenjak-Černe,
University of Ljubljana, Faculty of Economics, e-mail: simona.cerne@ef.uni-lj.si

Nataša Kejžar,
University of Ljubljana, Faculty of Social Sciences e-mail: natasa.kejzar@fdv.uni-lj.si

Vladimir Batagelj,
University of Ljubljana, Faculty of Mathematics and Physics, e-mail: vladimir.batagelj@fmf.uni-lj.si

model, for each country each age group can be considered as a separate variable for each sex. In this model, each country is presented with 34 variables: 17 variables for 5-years age groups for men, and 17 variables for 5-years age groups for women. Values are normalized so that they present percentages of the country's population in each age group. Euclidean distance between corresponding vectors is used. Although some objections against the usage of this dissimilarity measure can be found (Andreev and Andreev, 2004), by our opinion in comparing of the shapes of the population pyramids, each age-group can be considered as a separate variable.

Applying standard Ward's hierarchical clustering method (Kaufman and Rousseeuw, 1990) on pyramids, the corresponding 'centroid' pyramids of the clusters of countries are not real population pyramids describing the whole population in the clusters (Korenjak-Černe et al., 2008). So they can be interpreted only in terms of shapes, not as a population pyramids of countries' clusters.

In this paper we consider an alternative approach. The population pyramids can also be described with two vectors (one for each gender) of frequencies, representing the number of men/women for each age group. They form so called histogram symbolic data (Billard and Diday, 2006). For clustering so represented population pyramids the adapted Ward's hierarchical clustering procedure is used which was implemented in R (R Development Core Team, 2008). We will present the clustering results for the population pyramids of world countries and compare them with the results obtained using the traditional approach.

References

Andreev L., Andreev M. (2004). Analysis of Population Pyramids by a New Method for Intelligent Pattern Recognition. *Matrixreasonong*, Equicom, Inc..

Billard L., Diday E. (2006). *Symbolic data analysis. Conceptual statistics and data mining*. Wiley.

Kaufman L., Rousseeuw P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Korenjak-Černe S., Kejžar N., Batagelj V. (2008). Clustering of population pyramids. *Informatica*, Jun. 2008, 32 (2), 157-167.

IDB: International Data Base. http://www.census.gov/ipc/www/idbnew.html.

R Development Core Team(2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Modeling fertility and education in Italy: time-variant or invariant unobserved heterogeneity component?

Daniele Vignoli, Alessandra Mattei, Anna Gottard

1 Background and objective

The association between fertility and educational attainment is one of the most robust relationships recorded in social science. Education is in fact a potent marker of one's labour market performance and prospects, earnings potential, and social status. Overall, documented findings have shown that the direction of the effect of education on fertility depends on women's parity-specific status: better educated women have lower first birth intensities (e.g., Matysiak and Vignoli, 2009), whereas the effect of education on second order fertility is found to be positive in many European countries (e.g., Kreyenfeld, 2002; Kravdal, 2001).

Øystein Kravdal (2001) and Michaela Kreyenfeld (2002) gave a strong contribution to the debate suggesting the existence of a self-selection effect. They anticipated that some women with tertiary education who gave birth to the first child have a marked and unobserved preference for children. Following the methodological framework proposed by Lillard and Panis (2003), they tested this hypothesis employing a simultaneous-equations model that estimates jointly the transition to first and second birth and allows the inclusion of a person-specific heterogeneity term $\varepsilon_i \sim N(0, \sigma^2)$ is in each equation. Controlling for this unobserved component, that they interpreted as women's family-orientation, the positive effect of education on second birth risks vanishes . The major limitation of Kravdal's and Kreyenfeld's approach is that they considered a family-orientation constant over time, using a time-invariant individual-level unobserved-heterogeneity factor in modelling first and second birth transitions.

Daniele Vignoli,
Department of Statistics-University of Florence, e-mail: vignoli@ds.unifi.it

Alessandra Mattei,
Department of Statistics-University of Florence, e-mail: mattei@ds.unifi.it

Anna Gottard,
Department of Statistics-University of Florence, e-mail: gottard@ds.unifi.it

The objective of this paper is twofold. The role of educational attainment for fertility of Italian women will firstly be explored at the presence of time-invariant heterogeneity. This model can describe a persistent family orientation over the life-course. Secondly, the hypothesis of time-invariance will be relax to take into account for possible changes in family orientation during the life-course.

2 Data, method and preliminary results

The analysis is based on retrospective data on a sample of 5,506 women, stemming from the 2003 Istat Household Multipurpose Survey Family and Social Subjects.

We develop a full Bayesian framework, allowing us not only to account for the dependence of recurrent failure times by subject-specific random frailty, but also for the order of events, which is a crucial feature of dependence. We consider three model: the ordinary survival model for the transition to second conception (without random frailty component); the survival model including a person-specific random frailty component, assumed to be constant over time; and the survival model with a time-dependent random frailty, thus relaxing the time-invariant assumption.

Our preliminary findings show that after controlling for a common unobserved time-invariant and variant heterogeneity factor in each fertility transition, the significant (positive) impact of women's tertiary education vanishes. In this respect, the impact of women's tertiary education on Italian fertility development does not illustrate any significant impact when it is not amplified by those women who self-select themselves into family formation.

Further development of the study goes towards alternative specifications of the time-variant frailty component in order to better control for possible changes in women's family-orientation over time.

References

Kravdal O. (2001). The high fertility of college educated women in Norway: An artifact of the separate modelling of each partiy transition. *Demographic Research*, 5(6), 185-216

Kreyenfeld M. (2002). Time-squeeze, partner effect or selfselection? An investigation into the positive effect of women's education on second birth risks in West Germany. *Demographic Research*, 7, 15-48.

Ibrahim J.G., Chen M-H., Sinha S. (2001). *Bayesian Survival Analysis*, Springer Series in statistics, Springer.

Lillard L., Panis C.W.A. (2003). *aML Multilevel Multiprocess Statistical Software*. Release 2.0. EconWare, Los Angeles, California.

Matysiak A., Vignoli D. (2009). Finding the "right moment" for the first baby to come: A comparison between Italy and Poland. *MPIDR WP*, 2009-011.

Contributed Session 6

Handwriting and Web Analysis

The evaluation of handwriting evidence: multi-level models for determining authorship

Silvia Bozza, Franco Taroni, Raymond Marquis, Matthieu Schmittbuhl

1 Introduction

The evaluation of handwriting evidence in presence of questioned documents is an open problem in forensic science. The individualization of handwriting is still essentially based on the experience of experts, while Courts seem less and less at ease with experts opinions based solely on subjective belief. The contour shape of loops of characters can be studied in a global and quantitative way by following a methodology based on Fourier analysis (Marquis et al., 2005). This methodology allows a precise and objective reconstruction of characters' loops through a set of variables representing a set of harmonics. A methodology for the evaluation of handwriting evidence based on multivariate likelihood ratio was proposed in 2008 (Bozza et al., 2008). The methodology was based on a two-level Bayesian model, where the hierarchical ordering took into account the within- and the between-writers variability.

A kernel distribution is now proposed to replace the assumption of normality in the modellization of the distribution of between-group variability, and performances are compared. An additional level of variation will be added to take into account the different variability that characterize each specific letter (e.g., letters *a* or *o*) written by the same writer. The final model must allow for three sources of variation: (i) replicate measurements of the same character, (ii) between characters written by the same writer, (iii) between writers.

Silvia Bozza,
Dipartimento di Statistica, Università Ca' Foscari, Venezia, e-mail: silvia.bozza@unive.it

Franco Taroni,
School of Criminal Sciences, University of Lausanne, e-mail: franco.taroni@unil.ch

Raymond Marquis,
School of Criminal Sciences, University of Lausanne, e-mail: raymond.marquis@unil.ch

Matthieu Schmittbuhl,
Dental Faculty, University of Strasbourg, e-mail: matthieu.schmittbuhl@odonto-ulp.u-strasbg.fr

2 Data and models

The background data consist of p variables measured on each character in a sample of m writers, with n_i measurements on each writer, and are denoted as $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$, with $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$. Denote the mean vector within group i by $\boldsymbol{\theta}_i$ and the matrix of within group variances and covariances by W_i. Then, the distribution of $\mathbf{x}_{ij}$ is taken to be normal with $(\mathbf{X}_{ij} \mid \boldsymbol{\theta}_i, W_i) \sim N(\boldsymbol{\theta}_i, W_i)$. For the between-group variation, denote the mean vector between groups by $\boldsymbol{\mu}$ and the matrix of variances and covariances B. The distribution of the $\boldsymbol{\theta}_i$ can be taken to be normal with $(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, B) \sim N(\boldsymbol{\mu}, B)$. A Wishart inverse distribution is introduced to model the within-group variation, $W_i \sim IW(U, n_\omega)$. A multivariate normal distribution for $\boldsymbol{\theta}$ may not necessarily be a reasonable assumption, as it can be observed by inspection of bivariate plots of the means. The assumption of normality can be replaced by considering a kernel density estimate for the between-group distribution (Aitken and Lucy, 2004). The estimate of the overall density function is

$$\pi(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_m, B, h) = \frac{1}{m} \sum_{i=1}^{m} K(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_i, B, h), \tag{1}$$

where $K(\cdot)$ is the kernel density function that is taken to be a multivariate normal with mean at $\bar{\mathbf{x}}_i$ and covariance matrix $h^2 B$, and h is the smoothing parameter.

Suppose an anonymous letter, $\mathbf{y}_1$, and written material selected from a suspect, $\mathbf{y}_2$, are available for comparative purposes. The evaluation of handwriting evidence is performed through the derivation of a likelihood ratio, that provides a measure of the degree to which the evidence is capable of discriminating among propositions (i.e., $H_{1(2)}$: the suspect is (is not) the author of the manuscript):

$$LR = \frac{\int f(\mathbf{y}_1, \mathbf{y}_2 \mid \boldsymbol{\psi}, H_1) \pi(\boldsymbol{\psi} \mid H_1) d\boldsymbol{\psi}}{\int f(\mathbf{y}_1, \mathbf{y}_2 \mid \boldsymbol{\psi}, H_2) \pi(\boldsymbol{\psi} \mid H_2) d\boldsymbol{\psi}}, \tag{2}$$

where $\boldsymbol{\psi}$ is the vector of unknown parameters. A real case will be presented and discussed.

References

Aitken C., Lucy D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53, 109-122.

Bozza S., Taroni F., Marquis R., Schmittbuhl M. (2008). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Applied Statistics*, 57, 329-341.

Marquis R., Schmittbuhl M., Mazzella W., Taroni F. (2005). Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of capital characters 0. *Forensic Science International*, 150, 23-32.

Relative Linkage Disequilibrium in tracking web search patterns

Ron Kenett, Silvia Salini

1 Introduction

In this paper we apply association rules to data generated by the Google Insight application tracking frequency of keywords searched by various categories such as time and location (see http://www.google.com/insights/search/). Our analysis extends the Google Insight report by indicating relationships between pairs of searched keywords. We use Relative Linkage Disequilibrium (RLD) as a measure for characterising association rules. RLD was originally proposed as an approach to analyse quantitatively and graphically general two way contingency tables (Kenett, 1983). It has recently been extended to general association rules (Kenett and Salini, 2008a,b). As we will demonstrate, RLD can be interpreted graphically using a simplex representation leading to powerful graphical displays of association relationships. The idea is to show, with a simplex representation, the relationship between keywords over the time. Such relationships present interactions and interesting patterns, not revealed by univariate analysis.

2 Main results

In evaluating the structure of a 2x2 contingency table we consider four relative frequencies, x_1, x_2, x_3, x_4, $\sum_{i=1}^{4} x_i = 1, 0 \leq x_i, i = 1...4$.. The linkage disequilibrium is calculated as $D = x_1 x_4 - x_2 x_3$.

There is a natural one to one correspondence between the set of all possible 2x2 contingency tables and point on a simplex. We exploit this graphical representation to map out association between terms searched by Google in the last 6 years in

Ron Kenett,
KPA Group and University of Turin, e-mail: ron@kpa.co.il

Silvia Salini,
University of Milan e-mail: silvia.salini@unimi.it

USA, for example *Depression* and *Obama*. Table 1 show the frequencies for the terms *Depression* and *Obama* in the last six year in the USA, the values of D and its relative measure RLD that range fro 0 to 1. Figure 1 show on the 2D and 3D simplex this six contingency tables.

Table 1 Frequencies, D and RLD for *Depression* and *Obama* from 2004 to 2009 in USA

Year	Dep&Oba	Dep&ˆOba	ˆDep&Oba	ˆDep&ˆOba	D	RLD
2004	0,0049	476,9596	112,5412	10,4943	-0,1491	0,99995
2005	0,0168	433,2366	144,7836	21,9630	-0,1742	0,99984
2006	0,0156	464,2730	111,3604	24,3510	-0,1436	0,99982
2007	55,2900	242,0395	149,2769	153,3936	-0,0768	0,45459
2008	155,6830	134,3336	90,7036	219,2798	0,0610	0,28744
2009	179,5580	109,8548	84,0307	226,5565	0,0874	0,38414

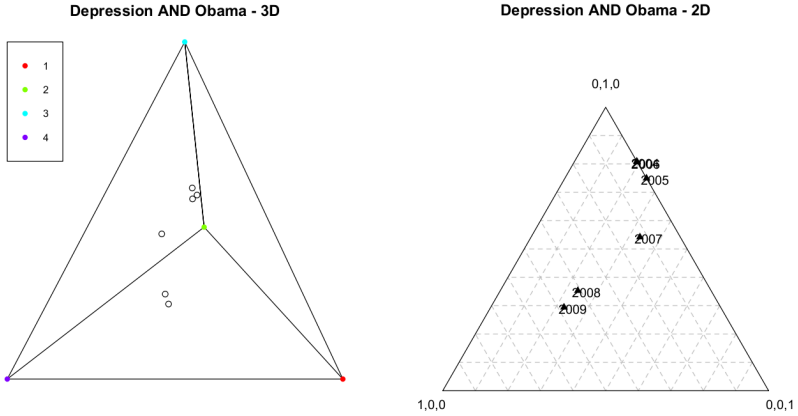


Fig. 1 Simplex representation of *Depression* and *Obama* in the last years in USA

References

Kenett R.S. (1983). On an Exploratory Analysis of Contingency Tables. *The Statistician*, 32, 395-403.

Kenett R.S., Salini S. (2008a). Relative Linkage Disequilibrium: A new measure for association rules. in: P. Perner (Ed.), Advances in Data Mining: Medial Applications, E-Commerce, Marketing, and Theoretical Aspects, ICDM 2008, Leipzig, Germany. Lecture Notes in Computer Science, Springer Verlag, 5077.

Kenett R.S., Salini S. (2008b). Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks. *Transactions on Machine Learning and Data Mining*,1 (2), 83-96.

Baysian classification tree for statistical user web-URL categories navigation pattern model

Klaokanlaya Silachan, Panjai Tantatsanawong, Chidchanok Lursinsap

1 Introduction

Users' Web access behavior model or navigation pattern is part of web usage mining. The focus here is the web navigation pattern from sets of URLs visited by users in access log, which helps to understand online users' need. The access log can categorize websites using URL as features, which is called Web-URL categories. The statistic and rules, focused in finding users' behaviors for Web-URL categories, can be used for pattern discovery. By using classification technique, we combine Bayesian and C4.5 to map data items into one of several predefined classes, which help to establish a profile of users belonging to a particular class or category on the Web-URL categories. This Bayesian classification is probabilistic learning to calculate explicit probabilities for hypothesis and is among the most practical approaches to certain types of learning problems. In this paper,the focus is to discover the rules. We use C4.5 as a decision tree; on which the statistic pattern can be shown; to build a model for classifying one attribute based on the other attributes. From the data sets, we can use a formula to obtain statistic from the users' behavior.The combining algorithm approach is not only useful to discover rules and interesting matters in users' Web access behavior model, but also helps to reduce steps in classifying and displaying pattern discovered from users' Web-URL categories access behavior (Mobasher et al., 2000; Quinlan, 1993; Silverstein et al., 1998; Buntine, 1992).

Klankanlaya Silachan,
Department of Science, Silpakorn University, Thailand,e-mail: klao_99@yahoo.com

Panjai Tantatsanawong Ph.D,
Department of Science, Silpakorn University, Thailand, e-mail: panjai@su.ac.th

Prof Chidchanok Lursinsap Ph.D,
Department of Matematics, Chulalongkorn University, Thailand, e-mail: lchidcha@hotmail.com

2 Main results

Baysian and C4.5 are combined to construct tree and statistic notation for trees and discovery rules for Web-URL categories,by calling Web-URL Bayesian categories tree. The new formula equation for studying this web user behavior is based on Bayesian principles, which comprises the number of user categories and the number of URL in each category, respectively.

Baysian (Buntine, 1992; Dunham, 2002) Bayes classification rule is the learning which used probability theory based on Bayes theorem. The goal is to build a model to find out the most probable hypothesis using probability. Given a set of data X = $\{x_1, x_2, x_3, ..., x_n\}$, a data mining problem is to uncover properties of the distribution from which the set comes. Bayes rule can be expressed as follow

$$P(h_1|x_i) = P(x_i|h_1)P(h1)/P(x_i|x_1)P(h1) + P(x_i|h_2)P(h_2) \quad (1)$$

is called the posterior probability, while P(h1) is the prior probability.$P(x_i)$ is the probability of the occurrence of data value and $P(x_i|h_1)$ is the conditional probability.

C4.5 (Quinlan, 1993) is an algorithm used to generate a decision tree. The attribute with the highest gain ratio is assumed to be the best for classification. Information gain depends on the probability of data. In order to calculate Information Gain(I) and entropy of attribute A (E(A)), Then, we calculate gain. The gain used to select attribute A to be a node of the tree has to be equal to the quantity of needed data, so that the data can be classified. Then, subtract it the gain with the quantity of needed data in order to classify data using attribute. Gain(A) = I($S_1, S_2, ..., S_m$) - E(A) while S_i/S_m denotes probability of data subset. Then, Gain Ratio = Gain Split Information. Finally, the highest Gain Ratio is chosen to be the root node, and the less Gain Ratio to be leaf nodes, respectively.

References

Dunham M. H. (2002). Data mining Introductory and Advanced Topics. *Chapter 3*, p.52.

Mobasher B., Cooley R., Srivastava J.(2000). *Automatic personalization based on web usage mining*, Communications of the ACM, vol.43, 142-151.

Quinlan J. R.(1993). Programs for Machine Learning, Morgan Kaufmann.

Silverstein C., Brin S., Motwani R., Ullman J. (1998). *Proceedings of the 24th VLDB conference*, New York, 594–605.

Buntine W.(1992). Learning Classification Trees. Statistics and Computing.

Contributed Session 7

Applications to Sociology and Market

Integrating the spatial dimension into propensity score matching to evaluate regional impact of capital subsidies

Marusca De Castris, Guido Pellegrini

1 Introduction

There are very few papers analyzing the spatial diffusion of industrial policies based on subsidies to private firms (De Castris and Pellegrini, 2008). The presence of spatial interactions implies that subsidies to firms in a region also affect firms in contiguous areas. The focus of our paper is to identify the employment growth effects of industrial subsidies when regional employment growth is spatially correlated. We propose a novel "spatial propensity score matching" technique that allow to correct for the spatial bias. The new estimator is applied on a new dataset that incorporates information on the two more important measures for local development in Italy: incentives to private capital accumulation by Law 488/92, mainly devoted to SME, and program agreement, created for large projects. The analysis is based on the grid of the local labour systems (LLSs) in Italy.

2 Main results

We want to measure the impact of an industrial policy program on employment (Y) in each LLS. Potential bias due to non random selectivity into the policy intervention is controlled by a non-experimental causality inference approach, using a propensity score matching estimator. The estimation of a "classical" propensity score in our empirical application shows that the residual are spatially correlated. The reason is the presence of a spatial crowding out: the subsidized firms could replace firms and

Marusca De Castris,
Dept. Public Institutions, Economy and Society, Roma Tre, e-mail: decastri@uniroma3.it

Guido Pellegrini,
Dept. Economic Theory and Quantitative Methods for Political Choices, Sapienza,
e-mail: guido.pellegrini@uniroma1.it

investments project in the neighbouring areas, by a spatial crowding out effect in the input, output and in the labour markets. In this case the probability of LLS to be treated (to have a substantial share of investments that is subsidized) depends on the probability that the contiguous LLSs are treated. From an econometric point of view, the problem can be deal with estimating an appropriate spatial lag model (Anselin, 1998) for the spatial propensity score (p_{spat}). A generic specification is $p_{spat} = F(h(X_i), g(W * p_{spat}))$, where $F(.)$ is the logistic cumulative distribution and $h(X_i)$ is a function of covariates, W is a spatial weights matrix where each element, w_{ij} , is equal to 1 if the distance between two centroids of LLS is less than the average distance between the centroids, 0 otherwise. The coefficient of $W * p_{spat}$ cannot be estimated by a simple logit model, given the effect of error term on $W * p_{spat}$. Our approach is to use an IV estimator, instrumenting the $W * p_{spat}$. Given a correct estimate of p_{spat} , a usual matching approach can identify the effects of industrial scheme on the outcome. We consider as subsidized areas the LLS having a share larger than 5% of subsidized new employment. We estimated a 2-stage probit model with endogenous regressors: at the first stage W* pspat is instrumented, using the exogenous covariates and their spatial lag. The result is plugged into the second stage equation. Using the spatial propensity score, we implemented the matching algorithm. in order to identify the average treatment effect on treated (ATT) with respect to employment dynamics. We use a Kernel and a Nearest Neighbour matching estimator; the standard errors of the ATT are estimated by the bootstrap procedure. The two estimated propensity score are used in order to detect the average effect of subsidies (Table 1). In both cases the effects are positive and statistically significant. However, adjusting for the spatial bias, the ATT is clearly larger.

Table 1 Effects of industrial subsidies on employment growth (1996-2001)

	Kernel Matching Estimation					Nearest Neighbour Matching Estimaton				
	Treated	Control	ATT	S.E.	t	Treated	Control	ATT	S.E.	t
Empl. Growth	190	283	0.026	0.013	2.02	190	107	0.04	0.019	2.08
Empl. Growth (spatial p.s.)	190	264	0.036	0.015	2.35	190	95	0.056	0.018	3.22

References

Anselin L. (1998). *Spatial Econometrics, Methods and Models*. Boston, MA: Kluwer Academic.

De Castris M. and Pellegrini G.,(2008). *Identification of the spatial effects of industrial subsidies* CREI Working papers, n. 4/08.

Asset ownership of the elderly across Europe: a multilevel latent class analysis to segment country and households

Omar Paccagnella and Roberta Varriale

1 Introduction

Wealth is an useful measure of the socio-economic status of the elderly, because it might reflect both accumulated socio-economic position and potential for current consumption. A growing number of papers have studied household portfolio in old age, both from a financial point of view and from a marketing perspective.

This paper aims at shedding more light on the behaviour of households across Europe by investigating similarities and differences in the ownership patterns of many financial and real assets. To this aim, we exploit the richness of information provided by SHARE (Survey of Health, Aging and Retirement in Europe), an international survey on aging that collects detailed information on several aspects of the socioeconomic condition of the European elderly (Börsch et al., 2008).

Overall, our sample contains 23238 households living in 14 countries (Sweden, Denmark, Ireland, Germany, The Netherlands, Belgium, France, Switzerland, Austria, Spain, Italy, Greece, Poland and Czech Republic). The focus is on the ownership of: bank or postal accounts; bonds; stocks; mutual funds; individual retirement accounts (IRAs); life insurance; house; mortgage; other real estate.

Given the hierarchical structure of the data, the econometric solution we adopt is the multilevel Latent Class (LC) model introduced by Vermunt (2003), that allows to obtain simultaneously country and household segments. The probability of observing a particular response pattern for country j is expressed by three components: (i) the probability that country j belongs to higher-level LC t; (ii) the probability that household i belongs to first-level LC k, given the country latent class membership;

Omar Paccagnella
Department of Statistical Sciences, University of Padua, via C. Battisti, 241/243 - 35121 Padova, Italy, e-mail: omar.paccagnella@unipd.it

Roberta Varriale
Department of Statstics "G.Parenti", University of Florence, Viale Morgagni, 59 - 50134 Firenze, Italy e-mail: roberta.varriale@ds.unifi.it

(iii) the probability that household i of country j owns the financial product h, given the household latent class membership.

2 Main results

The final model classifies the households in eight latent classes and the countries in six latent classes. The largest lower-lever LCs (number 5, 7 and 8, with the probability that household i belongs to one of them greater than 50%) are characterized by people with a low probability of holding any risky assets, such as stocks. Households with very large and diversified portfolios (in terms of probability of owning different financial and real assets) are in LC3 and LC4. LC2 is characterized by households with a high probability of owning the accommodation, a mortgage and bank accounts: the "oldest old" people typically do not have mortgages, so this class might include "younger" households (i.e. individuals who are still at work).

Using the empirical Bayes modal prediction, the countries may been assigned to one of the six latent classes. Central Europe countries (plus Ireland) have the highest probability to be grouped together and they are characterized by an interesting mix of household segments with non-risky (lower-level LC5 and LC7) and risky asset portfolio (lower-level LC3). According to the model, the richest countries (Sweden, Denmark and Switzerland) belong to a higher-level class where the probability that households belong to lower-level LCs with risky asset portfolio (LC3 and LC4) is the highest. Mediterranean countries are mainly characterized by household segments with a low probability of risky asset portfolio. Each of the remaining LCs are composed by a single country: the Netherlands, Poland and Czech Republic.

Our results confirm the evidence of a substantial heterogeneity of portfolio holdings of the elderly within and across Europe. However, the innovative features of our approach allow us to highlight two other important findings: first, in each country we can recognize some groups of holders whose portfolio characteristics are similar to those in other European countries (e.g. the richest households); second, when analysing household portfolios of the elderly in Europe, some countries are characterized by so similar patterns of asset ownership that can be grouped together (instead of specifying a full set of country-specific effects).

References

Börsch-Supan A., Brugiavini A., Jürges H., Kapteyn A., Mackenbach J., Siegrist J., Weber G. EDS (2008). *First Results from the Survey of Health, Ageing and Retirement in Europe (2004-2007). Starting the Longitudinal Dimension.* MEA, Mannheim.

Vermunt J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Restructuring and innovations on the survey "capacity of collective tourist accommodation" and their impact on the process quality

Maria Teresa Santoro, Simona Staffieri

1 The survey on capacity of collective tourist accommodation

The survey covers the supply of collective tourist accommodation in Italy, giving information on the available capacity. A collective tourist accommodation is an accommodation establishment that provides overnight lodging for travelers in a room or some other unit; it differs from a private establishment due the number of beds available needing to be greater than a specified minimum and all the places in the establishment must come under a common commercial-type management, even if it is non-profit-making. The survey is currently carried out by ISTAT as a yearly census carried out in the frame of the European Union Council Directive 95/57/CE of 23/11/95 and according to the national rules set by ISTAT. The survey unit is the collective accommodation establishment, classified in the following categories:

1. hotels and similar establishments (classified in 5 stars categories);
2. other collective accommodation establishments: camp-sites, holiday villages, camp-sites - holiday villages (mixed form), holiday dwellings, farmhouses country-houses, youth hostels, mountain refuges, other collective accommodations;
3. rented private tourist accommodation: B&B e others.

Variables of interest are those related to the capacity of tourism establishments and specifically: number of establishments, bed-places, rooms and bathrooms (for hotels and similar establishments); number of establishments, bed-places (for the other types of establishments). The data relates to a specific year and the variables regarding hotels and similar establishments are collected also by size classes (in terms of number of rooms) of this kind of units. The geographical breakdown of the data collection as well as of the data dissemination is the municipality. Data are

Maria Teresa Santoro,
Istat - Division for Short-Term economic statistics, e-mail: masantor@istat.it

Simona Staffieri,
Istat - Division for Short-Term economic statistics, e-mail: staffier@istat.it

provided by the local bodies in charge for tourism, that fill an electronic questionnaire with internal checks (the so-called Mod.ISTAT CTT/4). Data are sent by web. Mod.ISTAT CTT/4 collects the capacity variables, measuring the situation of the national tourism accommodation supply in each calendar year, recording information independently from the seasonal occupancy (and closure of the establishments). So it records the gross capacity.

2 Restructuring and innovations on the survey

During the year 2009, the survey was restructured. A new data-base was designed. The intermediate bodies in charge of tourism statistics (Regions and Provinces) were requested to send the electronic questionnaire using the web device INDATA (ISTAT certified on -line site of data capturing). The electronic questionnaire already contains the data of the previous year. Respondents have only to update changes. This practice implies a reduction of statistical burden. Once the model is completed, a control procedure allows the verification of the data correctness and the consistency of changes over time in the crucial variables. Data are uploaded in a space dedicated to tourism in a ISTAT's server. The new system SITCARI, created to manage data on capacity of collective tourist accommodation, allows to:

1. upload models and check models;
2. monitor the status of the model (arrived/not arrived; recalled/not recalled; the kind of recall: mail, telephone, letter);
3. download micro-data quality reports (list of warning, errors and changes);
4. download tables of statistical thresholds to data dissemination;
5. predispose tables for data dissemination.

The information on the status and the reports of the process of quality checks allow to define various profiles of the respondents (regarding the performance with respect to many dimension: timeliness on sending data, quality and coverage of data, types of errors). In particular, those profiles are defined applying a Cluster Analysis approach, using a hierarchical algorithm . Moreover it is analyzed what kind of recall provides the best results in terms of effectiveness.

References

Eurostat (1998). *Methodological Definitions to be Taken into Account when Collecting Statistical Information on Tourism as requested in the Council Directive n. 95/57/EC of 23 November 1995*, Lussemburgo, marzo 1998.

Eurostat (2008). *International recommendations for tourism statistics*, Madrid 2008.

Santoro, M.T (2005). Il quadro europeo delle statistiche sul turismo. In: XIV Rapporto sul turismo italiano 2005, edizioni Mercuri, Roma, 27 settembre 2005.

Contributed Session 8

Discrete Data

Generating ordinal data

Pier Alda Ferrari, Alessandro Barbiero

1 Introduction

In the recent years, a great interest has been devoted by researchers to categorical data and the related statistical methods employed for their joint analysis. Specifically in explorative analysis, the robustness and performance of these techniques can be assessed almost exclusively through simulation studies, which require to generate a huge number of datasets, according to some experimental conditions.

A general method for obtaining data with a desired pattern is proposed by Cario and Nelson (1997) and it is called NORTA (NORmal To Anything). This method produces random vectors with fixed marginal distributions and correlation matrix starting from a standard multivariate normal. This method has been extended by Stanhope (2004), trying to overcome some practical drawbacks. With regard, more properly, to ordinal data, Demirtas (2009) proposes a method for generating multivariate ordinal data given marginal distribution and correlation matrix R^{ORD*}. His technique first generates binary data by collapsing the corresponding ordinal categories and then, through an iterative procedure, finds a proper binary correlation matrix R^{BIN}, which assures for the ordinal data the desired correlation structure. Even if very flexible, this method presents some limits. In this paper the focus is on ordinal variables and a simple procedure to obtain multivariate ordinal variables with specified marginal distributions and correlation structure, no longer impaired by previous drawbacks, is proposed and its performance is investigated through a simulation study and two applications.

Pier Alda Ferrari,
Department of Economics, Business and Statistics, Università degli Studi di Milano, Italy,
e-mail: pieralda.ferrari@unimi.it

Alessandro Barbiero,
Department of Economics, Business and Statistics, Università degli Studi di Milano, Italy
e-mail: alessandro.barbiero@unimi.it

2 Main results

Let Y_j be an ordinal variable, with k_j ordered categories, which is labeled in ascending order with the first k_j integer numbers. The problem is how to generate m ordinal variables, Y_1, Y_2,..., Y_m, components of a multivariate variable Y, such that $P(Y_j = k) = p_{jk}$ and $corr(Y_i, Y_j) = \rho_{ij}^{ORD*}$, where p_{jk}, $j = 1,2,\ldots,m$; $k = 1,2,\ldots,k_j$ and ρ_{ij}^{ORD*}, $i = 1,\ldots,m-1$, $i < j \leq m$ are specified.

The method here proposed, which is implemented in R, works as follows:

1. (setting up experimental conditions) Starting from $N(\mathbf{0}, R^{CONT} = R^{ORD*})$, we obtain, by a straightforward discretization process, which employs a quantile approach, ordinal data with assigned marginal distributions. We compute the new R^{ORD}, compare it with R^{ORD*} and adjust the R^{CONT} through an iterative procedure until R^{ORD} converges to R^{ORD*}; we also check out at each iteration if R^{CONT} is actually definite positive and if it is not we fix it through a "nearest correlation matrix" procedure, see Higham (2002). The final R^{CONT} is denoted as R^{CONT*}.
2. (generating data) The ordinal data are got by drawing a sample from $N(\mathbf{0}, R^{CONT} = R^{CONT*})$ and following the discretization procedure described above.

A preliminary simulation study has assured the good performance of this method under different scenarios. Then, the method has been applied to compare the performance of different missing data treatments on ordinal data under different experimental conditions (i.e. different number of categories and probabilities for each variable, and different association structure). Similarly, it has been used to compare the performance of NLPCA versus PCA for ordinal data, varying again the experimental conditions.

References

Cario M.C., Nelson B.L. (1997) Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois

Demirtas H. (2009) A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, **76**(11), 1017-1025

Higham N. (2002) Computing the nearest correlation matrix - a problem from finance; *Journal of Numerical Analysis*, **22**, 329-343

Stanhope S. (2004) Correlation control in small-sample Monte Carlo type simulations I: A simulated annealing approach, *Insurance: Mathematics and Economics*, **1**(1), 68-79

Gini measure: its decomposition proposal in the discrete case

Paolo Giudici, Emanuela Raffinetti

1 Introduction

The Gini measure decomposition always assumed the role of partioning the total inequality of a population into two components concerning the inequality between and the within subpopulations: Dagum (1997) suggested to decompose the Gini ratio into three components taking into account also the contribution of the intensity of transvariation between subpopulations. Our research aim consists in proposing a new approach devoted to define the Gini measure decomposition in terms of concordance and discordance: in order to achieve this goal one recurs to some specific statistical tools such as the concordance curve (see e.g. Giudici and Raffinetti (2009)). Let us suppose to have a k-variate random vector $(Y, X_1, \ldots, X_{k-1})$ and let us describe the relationship among the response variable Y and the explanatory variables $X_1, \ldots, X_{k-1}$ through the linear regression model application. Once building the response variable Lorenz curve and its dual, one proceeds to the concordance curve construction defined as the set of ordered pairs $(i/n, (1/nM_Y)\sum_{j=1}^{i} y_j^*)$ where $i = 1, \ldots, n$ and M_Y is the variable Y mean. In particular y_i^* represents the Y variable values ordered according to the ranks assigned to their respective estimates: we denote the concordance curve with $C(Y|r(\hat{y}_i))$ which moves between the Y Lorenz curve and its dual. Exploiting the perfect relationship between the Gini measure and the area between the Y Lorenz curve and its dual, we introduce a new Gini measure decomposition in terms of concordance and discordance: more precisely one can define the share of the Gini measure which corresponds to a concordance or to a discordance situation giving information about the Y variable concentration repre-

Paolo Giudici,
Statistics and Applied Economics Department, University of Pavia, Strada Nuova 65, 27100 Pavia (Italy), e-mail: giudici@unipv.it

Emanuela Raffinetti,
Statistics and Applied Economics Department, University of Pavia, Strada Nuova 65, 27100 Pavia (Italy), e-mail: emanuela.raffinetti@unipv.it

sented by the explanatory variables according to a ranks-based approach (see e.g. Giudici and Raffinetti (2009)). Let us define as concordance area, the area between the egalitarian line and the Lorenz curve, and as discordance area the area between the egalitarian line and the dual Lorenz curve. This kind of proceeding implies the study of a new form of dependence that we call *rank dependence* whose associated measure will be called *Gini rank dependence* (*GRD*).

2 Main results

Our proposal can be explained considering three different cases, as follows.
Case 1: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the concordance area.
Case 2: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the discordance area.
Case 3: the concordance curve $C(Y|r(\hat{y}_i))$ partially lies in the concordance area and partially in the discordance area meaning that between the concordance curve and the egalitarian line there is one or more intersections points. A general formulation of *GRD* is provided taking into account the different nature (even or odd) of the number of intersection points and the first segment position of the concordance curve with respect to the concordance/discordance area. Before proceeding we have to define some conditions: $a_j = 1$ with $j = p+1$ and $a_{j-1} = 0$ with $j = 1$, where a_j represents the intersection points x-axis values, with $j = 1, \ldots, p$. Furthermore if $s = 0 \Rightarrow$ the concordance curve initial position is in the discordance area, otherwise, if $s = 1 \Rightarrow$ concordance curve initial position is in the concordance area. The general Gini rank dependence measure equals to

$$GRD = (-1)^s \left\{ \sum_{j=1}^{p+1} (-1)^{j+1} \left[\int_{a_{j-1}}^{a_j} C(Y|r(\hat{y}_i))dy - a_j^2 \right] + \frac{1}{2} \right\}, \text{ if } p \text{ is even} \quad (1)$$

$$GRD = (-1)^{s+p} \left[\sum_{j=1}^{p+1} (-1)^{j+1} \left(a_j^2 - \int_{a_{j-1}}^{a_j} C(Y|r(\hat{y}_i))dy \right) \right] - \frac{1}{2}(-1)^s, \text{ if } p \text{ is odd} . \quad (2)$$

References

Dagum C. (1997). A new Approach to the Decomposition of the Gini Income Inequality Ratio. *Empirical Economics*, 22, 515-531.

Giudici P., Raffinetti E. (2009). Multivariate Ranks-based concordance indexes. In: Statistical Methods for the analysis of large data-sets 2009, Book of Short Papers, Meeting of the Italian Statistical Society, 439-442.

Adaptive discrete Beta kernel graduation of demographic data

Angelo Mazza, Antonio Punzo

1 Introduction

Demographic phenomena are often strongly related to age, in the sense that the intensity of events varies sharply across the age range.Therefore, a key aspect of demographic research consists in studying such age-specific patterns in order to discover regularities and make comparisons across time and space. Although in the following we will only refer to mortality, our approach can easily be extended to other age-dependent phenomena.

Let X be the variable age, with finite support $\mathscr{X} = \{0, 1, \ldots, \omega\}$, being ω the maximum age of death. We consider X as discrete, although age is in principle a continuous variable, since in demographic and actuarial applications age at last birthday is generally used; furthermore, discrete are also the commonly used age-specific indicators, such as the *proportion dying* q_x which is the proportion of persons who die at age x, $x \in \mathscr{X}$, and the *number dying* d_x which is the number of persons that would die at age x if starting from an arbitrary large hypothetical cohort l_0.

The use of such age-specific indicators is often deemed as not appropriate since a specific observed pattern may be often intended as a single realization of a stochastic phenomenon with certain distinctive mortality traits. Such random fluctuations are more of concern in actuarial studies and in applied demography, when small area datasets are investigated or when the events under investigation are particularly rare. To cope with this issue, several graduation techniques have been proposed in literature (see, *e.g.*, Debón et al., 2005, 2006, for an exhaustive comparison of parametric and nonparametric methods, respectively, in the graduation of mortality data), based on the assumption that if the number of individuals in the group on whose experience data are based were considerably larger, then the set of observed indicators would display a much more regular progression with age (Copas, 1983).

The most popular statistical method for nonparametric graduation is the kernel smoothing; in such method, most of the attention is usually dedicated to the selection of the smoothing parameter while symmetric kernel functions are routinely used.

Angelo Mazza, Antonio Punzo,
Dipartimento di Impresa, Culture e Società, Università di Catania (Italy),
e-mail: a.mazza@unict.it, antonio.punzo@unict.it

Nevertheless, if the use of symmetric kernels is appropriate when fitting functions with unbounded supports from both sides, its use is not adequate with age-dependent functions (Chen, 2000). When smoothing is made near the boundaries, in fact, fixed symmetric kernels do allocate weight outside the support (e.g. negative or unrealistic high ages) causing the well-known problem of boundary bias. Motivated by these considerations, in Mazza and Punzo (2010) is proposed a discrete kernel smooth estimator specifically conceived for the graduation of discrete finite functions. This approach, by construction, overcomes the problem of boundary bias, given that the kernels are chosen from a family of conveniently discretized and reparameterized beta densities whose support $\mathscr{X}$ matches the age range.

2 Main results

In this paper, we generalize the discrete beta kernel estimator proposed in Mazza and Punzo (2010) by making the bandwidth variable across the age range, according to the reliability of the data. So, for ages in which the amount of exposed e_x is higher, a lower bandwidth is used, resulting in an estimate closer to the observed rates.

We evaluate the performance of the proposed estimator through simulations. In detail, we resample from a model that applies a life table $\widetilde{q}_x$ to a population age structure e_x. For each sample, we compute the adaptive graduated $\widehat{q}_x$ and we measure the discrepancy between q_x and $\widehat{q}_x$. We have repeated the simulations according to different hypotheses of model life table, population age structure and population size. Comparisons with the (fixed) discrete beta kernel estimator proposed in Mazza and Punzo (2010) are also provided.

References

Debón A., Montes F., Sala R. (2006). A Comparison of Nonparametric Methods in the Graduation of Mortality: Application to Data from the Valencia Region (Spain). *International Statistical Review*, 74(2), 215–233.

Debón A., Montes F., Sala R. (2005). A Comparison of Parametric Models for Mortality Graduation. Application to Mortality Data for the Valencia Region (Spain). *SORT*, 29(2), 269–288.

Mazza A., Punzo A. (2010). Discrete Beta Kernel Graduation of Age-Specific Demographic Indicators. In: Ingrassia, S. and Rocci, R. and Vichi, M. EDS, New Perspectives in Statistical Modeling and Data Analysis, (in press), Berlin-Heidelberg-New York:Springer.

Chen S. X. (2000). Beta Kernel Smoothers for Regression Curves. *Statistica Sinica*, 10(1), 73–91.

Copas J. B. (1983). On the Choice of Bandwidth for Kernel Graduation. *Journal of the Institute of Actuaries*, 110, 135–156.

Contributed Session 9

Classification Methods

Restricted Boltzmann machines for market basket analysis

Harald Hruschka

1 Introduction

The market basket of a shopper consists of those product categories which s/he purchases during a shopping trip at a store. In this paper we look at purchase incidences and represent the market basket as binary vector whose dimension equals the number of considered categories.

The most prominent models capable to measure cross category effects for purchase incidence data are the multivariate logit (MVL) and the multivariate probit (MVP) models. The model we introduce is known as restricted Boltzmann machine (RBM) which differs from the MVL model by including binary hidden variables. So far the RBM has been mostly applied to solve pattern recognition problems (Hinton and Salakhutdinov, 2006).

Our effort can be characterized by two aspects. (1) We are able to deal with a high number of product categories, whereas the overwhelming majority of previous studies on multicategory purchase incidence using MVL or MVP models consider only between 4 and 12 numbers of categories (Russell and Petersen, 2006). (2) We develop and apply a simultaneous estimation method and do not need to follow a two step approach in contrast to Hruschka, Lukanowicz and Buchta (1999) or Boztuğ and Reutterer (2008) although we analyze a large number of categories

2 Main results

20,000 market baskets collected at a medium-sized supermarket during four consecutive Saturdays are randomly divided into two data sets of equal size. One set (estimation data) is required for estimation, the second set (validation data) is used

Harald Hruschka,
University of Regensburg, e-mail: harald.hruschka@wiwi.uni-regensburg.de

to determine the predictive accuracy of models. Out of a total of 209 categories, we analyze the 60 categories with the highest purchase frequencies.

We estimate the RBM by maximum likelihood estimation because factorizing over categories makes computation of the normalization constant practicable if the number of hidden variables is much lower than the number of categories. All the RBMs considered perform better than the independence model in terms of log likelihood values both for the estimation and the validation data. This result clearly shows that cross effects between categories should not be ignored. Log likelihood values of RBMs increase with the number of hidden variables. RBMs turn out to be very robust, as log likelihood values almost stay the same for the validation data.

We measure the marginal cross effect of category l on category j by the first derivative of the share of category j w.r.t. the share of category l (the share of a category j is the average value of purchase incidences y_{ij} across baskets) based on estimated shares of categories and estimated average values of hidden variables.

Tropical fruits and vegetables turn out to be the central product categories in terms of marginal cross effects. The cross effects that these two categories exert on each other are much greater in absolute values than all the remaining cross effects. Both tropical fruits and vegetables affect the same eleven categories, most strongly cheese, milk, water and butter. Among these eleven categories only one, beer, is subject to negative effects. Only one category, fruits, has (positive) effects on both tropical fruits and vegetables. Except for tropical fruits and vegetables asymmetry of marginal cross effects is pronounced. All cross effects of tropical fruits and vegetables on the eleven categories are asymmetric (e.g., tropical fruits and vegetables exert a significant effect on cheese, but cheese has no significant effect on tropical fruits and vegetables). Also, we found no significant effects of both tropical fruits and vegetables on fruits.

References

Boztuğ Y., Reutterer T. (2008). A Combined Approach for Segment-specific Market Basket Analysis. *European Journal of Operational Research*, 187 (1), 294-312.

Hinton G.E, Salakhutdinov R.R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313 (5786), 504-507.

Hruschka H., Lukanowicz M., Buchta C. (1999). Cross-Category Sales Promotion Effects. *Journal of Retailing and Consumer Services*, 6(2), 99-105.

Russell G.J., Petersen A. (2000). Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing*, 76 (3), 369-392.

Classification of chunked data using Proximal Vector Machines and Singular Value Decomposition

Antonio Irpino, Mario Rosario Guarracino, Rosanna Verde

1 Introduction

Classification refers to the capability of a system to learn from examples how to discriminate cases in two or more given classes. The system learns from a set of cases, usually referred as the *training set*. Each case is described by a set of features and the class label. For each new case, the trained system predicts its class label. In case of only two classes, the problem is called *binary classification*; in all other cases it is named *n-class* or *multiclass classification*. Support Vector Machines (SVM) (Vapnik, 1995) represent state of the art in supervised learning. Recently, the Regularized Generalized Eigenvalue Classifier (ReGEC) (Guarracino et al., 2007) extension has been proposed to solve binary and multiclass classification problems (Irpino et al., 2010). Nowadays, data production grows at an increasing rate, so it is common in different field of research to deal with huge amount of data. Further, with the explosion of networks of sensors, the data production is done collecting chunk of data and then processing them. A plausible solution for analysing such amount of data is the development of incremental or distributed algorithms. In the classification environment, this kinds of algorithms furnish partial models of classification that subsequently need to be fused (Sinha et al., 2008). When models of classification are based on linear functions, for separating (like in SVM) or representing (like in ReGEC) classes, we propose Singular Value Decomposition strategy for merging such models. Let Ω be a dataset of N labeled data (i.e. classified into $k > 1$ classes),

Antonio Irpino,
Dip. di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli,
e-mail: antonio.irpino@unina2.it

Mario Rosario Guarracino,
ICAR-CNR, Napoli e-mail: mario.guarracino@cnr.it

Rosanna Verde,
Dip. di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli,
e-mail: rosanna.verde@unina2.it

partitioned into r chunks, each one of cardinality N_r, and described by p explicative variables. We propose to define a classification model for each class in each chunk. In order to discover a unique classification model, we propose to use the Singular Value Decomposition of the linear models defined for each chunk and each class. This can be considered as a strategy for fusing classifiers.
In particular, we show the computational advantages of such strategy: it needs few computational resources and reaches a similar classification performance (in terms of accuracy) than classic algorithms in defining the unique classification model.

References

Vapnik V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag.

Guarracino M. R., Cifarelli C., Seref O., Pardalos P. (2007). A classification algorithm based on generalized eigenvalue problems, *Optimization Methods and Software*, 22(1), 73–81.

Irpino A., Guarracino M. R., Verde R. (2010). Multiclass Generalized Eigenvalue Proximal Support Vector Machines. 4th IEEE Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2010), February 2010, 25–32.

Sinha A., Chen H., Danu D. G., Kirubarajan T., Farooq, M. (2008). Estimation and decision fusion: A survey. *Neurocomputing* 71, 13–15 (Aug. 2008), 2650-2656. DOI= http://dx.doi.org/10.1016/j.neucom.2007.06.016

Robust kernel Fisher discriminant analysis with weighted kernels

Nelmarie Louw

1 Introduction

Kernel Fisher discriminant analysis (KFDA) is a kernel based extension of linear discriminant analysis (LDA), which was proposed by Mika et al. (1999). The KFD classifier is given by

$$sign\left\{b+\sum_{i=1}^{n}\tilde{\alpha}_i K(\mathbf{x}_i,\mathbf{x})\right\}. \tag{1}$$

Here $K(.)$ denotes a kernel function and the quantities b and $\tilde{\alpha}_1,\tilde{\alpha}_2,\cdots,\tilde{\alpha}_n$ are determined by applying the KFDA algorithm to the training data. One of the most popular kernels is the Gaussian kernel, defined by $K_G(\mathbf{x}_j,\mathbf{x}_k)=exp(-\gamma\|\mathbf{x}_j-\mathbf{x}_k\|^2)$, where γ is a so-called kernel hyperparameter that has to be specified beforehand or determined from the data. Although the KFD classifier performs well in many classification problems, its performance is adversely affected by the presence of outliers and noise. The aim in this paper is to develop a more robust KFD classifier. When there are outliers in the data it will result in an unusually large value of $\|\mathbf{x}_j-\mathbf{x}_k\|^2$, and thus an unusually small value of $K_G(\mathbf{x}_j,\mathbf{x}_k)$. Since the KFD algorithm operates on the entries in the kernel matrix, limiting the effect of outliers on these entries may reduce their effect on the KFD classifier. Our proposal is to use weight functions to achieve this. We propose the weighted kernel

$$K_w(\mathbf{x}_j,\mathbf{x}_k)=exp(-\gamma\sum_{i=1}^{p}w_i(x_{ji}-x_{ki})^2)=exp(-\gamma\|W^{1/2}(\mathbf{x}_j-\mathbf{x}_k\|^2), \tag{2}$$

with $W=diag(w_i)$. The weights are determined from the data, and we consider different ways of calculating the weights. Writing $z_i=|x_{ji}-x_{ki}|$, we propose using the weight functions that are often used in robust methods: $w_{1i}=(1+1/2(z_i/\delta)^2)^{-1}$

Nelmarie Louw,
Stellenbosch University, e-mail: nlouw@sun.ac.za

and $w_{2i} = 1$ for $z_i \leq \delta$; $w_{2i} = (\delta/z_i)^2$ for $z_i > \delta$. The value of δ has to be specified or determined from the data (through crossvalidation).

2 Main results

To evaluate the performance of the proposed weighted kernels, a detailed Monte Carlo simulation study was done. We consider training data generated from normal (denoted by N) and lognormal (L) distributions. We investigate cases where the populations differ with respect to location (LOC) and w.r.t. spread (SPR), and consider various covariance structures. We consider "uncontaminated" data with and without noise variables as well as data which were contaminated with 10%, 20% and 30% outliers. The Gaussian and weighted kernels were used to construct KFD classifiers, which were then used to classify large independent test data sets. Table 1 contains a representative selection of the resulting estimated error rates.

Table 1 Estimated error rates

		0	0.1	0.2	0.3
N-LOC	K_G	0.165	0.174	0.175	0.178
	K_{w_1}	0.105	0.123	0.128	0.142
	K_{w_2}	0.094	0.104	0.122	0.121
N-SPR	K_G	0.255	0.313	0.358	0.416
	K_{w_1}	0.134	0.220	0.275	0.408
	K_{w_2}	0.108	0.137	0.138	0.378
L-LOC	K_G	0.060	0.096	0.110	0.140
	K_{w_1}	0.046	0.059	0.086	0.098
	K_{w_2}	0.026	0.027	0.042	0.068
L-SPR	K_G	0.179	0.277	0.424	0.547
	K_{w_1}	0.062	0.174	0.241	0.309
	K_{w_2}	0.077	0.159	0.226	0.259

The KFD classifiers with the weighted kernels consistently yielded lower error rates than the KFD classifier with the Gaussian kernel. Similar results were obtained for the other configurations studied. We therefore conclude that the proposed weighted kernels successfully reduce the error rate of the KFD classifier.

References

Mika S., Rätsch G., Weston J., Schölkopf B., Müller K. R. (1999). Fisher discriminant analysis with kernels. In: Y.-H. Hu, J. Larsen, E. Wilson and S. Douglas, editors, *Neural Networks for Signal Processing*, 41–48.

Contributed Session 10

Time Series and Spatial Analysis

Forecasting and clustering beanplot time series

Carlo Drago and Germana Scepi

1 Introduction

Scalar Time Series Forecasting is the application of a statistical model to estimate and forecast the future values of a time series. There are real cases in which scalar time series x_t do not permit to correctly visualize and forecast the temporal event, as when the dataset contains an huge quantity of observations. This is the case, in Finance for example, of data collected at a given high frequency (for examples minutes), but analyzed at a lower frequency (daily). Here, there is a specific need of data aggregation. In the framework of Symbolic Data Analysis, it was proposed an alternative way to manage huge temporal datasets: transforming the original data in symbolic one as Intervals, Histograms and so on (Arroyo and Mate, 2006). Afterwards, we have shown the usefulness of beanplot time series (Drago and Scepi, 2009).

The beanplot can be considered as a particular case of interval-valued modal variable at the same time than boxplots and histograms. In a beanplot we can take in account the interval between the minimum and the maximum value, we can define a central measure and, principally, we can compute a density function by means of a kernel nonparametric estimator (Kampstra, 2008). Therefore the beanplot allows us to synthesize the location, the size and the shape of our data. Moreover, the beanplot allows us to visualize the "bumps", representing the values of maximum density and showing important equilibrium values reached in each temporal interval. A Time Series Beanplot $\{b_{x_t}\} t = 1...T$ is an ordered sequence of beanplots or densities over the time. The choice of the length of the temporal interval (day, month, year) depends both on the specific data features and on the objectives of the analyst. With a visualization aim, we use the Sheather-Jones criteria that defines the optimal interval

Carlo Drago,
University of Naples "Federico II", e-mail: carlo.drago@unina.it

Germana Scepi,
University of Naples "Federico II", e-mail: germana.scepi@unina.it

in a data-driven approach. In this paper, firstly we propose a beanplot forecasting approach based on a parametrization of these complex data considering beanplot time series as single ones. Secondly, dealing with multivariate beanplot time series, we propose to define homogeneous groups of time series for explorative aims.

2 Main results

The initial step of our strategy consists in developing an algorithm (in R) for the parameterization of the beanplot time series. The time series attributes are realizations of the single features of the beanplot, as the coordinates $\bar{y}$, $\bar{x}$ representing the density values and the initial values respectively. We refer to them as descriptor parameters of the beanplot structure. They allow us to define an attribute time series. In that sense we measure the evolutionary dynamics of the beanplots over the time. Successively, we have developed a forecasting algorithm for each time series of attributes by using a Vector Autoregressive Model. The gain in using a forecasting method on these types of data, is related to the possibility to predict their (and so the underlying financial data) complex behaviour over the time. For example it could be very important to detect and predict the increasing difference between minimum and maximum in periods with higher volatility.

Considering multivariate beanplot time series, we have classified their attributes using classical methods. The obtained results show different classes of behaviours (or response to shocks). In particular the results show that beanplots with a similar underlying model have similar behaviours. Anyway it is very interesting to note the deviations from this empirical rule, useful for example in statistical arbitrage. This clustering approach can be usefully applied to financial problems as, for example, the asset allocation with the aim of diversifying the portfolios.

References

Arroyo J., Mate C., (2006). Introducing Interval Time Series: Accuracy Measures. in COMPSTAT 2006, *Proceedings in Computational Statistics*, Heidelberg, 1139-1146, Physica-Verlag.

Drago C., Scepi G.,(2009). Univariate and Multivariate Tools for Visualizing Financial Time Series. In Classification and Data Analysis 2009, Book of Short Paper, 481-485.

Kampstra P., (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions *Journal of Statistical Software*, 28, Code Snippet 1, Nov 2008

Local analysis of spatial relationships: a comparison of GWR and OLS method

Massimo Mucciardi, Pietro Bertuccelli

1 Introduction

In recent years, the need and the interest of the scientific community for other methods to analyze and represent data has driven many researchers to explore new ways. One of the most interesting techniques provided by spatial statistical literature was the introduction by Brundson et al. (1996) of the Geographically Weighted Regression (GWR). In the light of this, we decided to conduct a more detailed analysis of the relationships between the TFR (Total Fertility Rate) of the 103 Italian provinces in 2005 and some well known social and demographic determinants based on the ordinary least squares (OLS) regression and the GWR method. The aim of this work is to show the improvement in model performance of GWR over OLS.

2 Main results

The variables chosen in the first stage were used for the construction of the local model. For the GWR model we have considered: 1) the provinces centroids for the distance calculation between spatial units; 2) the cross-validation criterion for the bandwidth selection 3) the AICc and the RSS for models evaluation; 4) the Leung test to reveal the spatial non-stationarity. Among the many estimated models for the explanation of provincial TFR, we chose a GWR model with an adaptive Gaussian kernel (Fotheringham et al., 2002). Finally, the determinants are showed in table 1. The analysis and the interpretation of the GWR estimates is done by keeping in count the global model (OLS) and the tests for spatial non-stationarity.

Massimo Mucciardi,
University of Messina, e-mail: mucciard@unime.it

Pietro Bertuccelli,
University of Messina e-mail: pbertuccelli@unime.it

The analysis of the OLS indicates that the TFR exhibits a statistically significant negative relationship with MACM (Mean age at childbearing for mother) and IMR (Internal migration rate) and statistically significant positive with SFW (Share of foreign women), IECM (Indirect expense for childbearing and maternity per capita) and MAR (Marriage rate for 1000 inhabitants); CMR (Civil marriage rate) isn't globally statistically significant but it shows a strong spatial non-stationarity. The results obtained indicate that local model significantly improved the OLS results with AIC value dropping from to -227.28 to -245.50 and R^2 rising from 0.61 to 0.77 (see tab.1). Moreover, the use of GWR model is necessary because the analysis on the OLS residuals reveals a strong spatial autocorrelation (see the Moran index for OLS (I_{OLSres}) and GWR (I_{GWRres}) residuals). Although there exists, at the national level, a negative highly significant relationship between the MACM and TFR, the GWR results make it clear that this relationship is highly variable in space. It may be observed (figure not shown) that the inverse relationship between MACM and TFR is stronger in the south of Italy and less elsewhere. In conclusion, our results confirm that the advantage of GWR over OLS is mainly due to the consideration of the true spatial variation of the relationship between fertility and socio-demographic determinants.

Table 1 Mean values of OLS and GWR parameters

Variable	Global Model (OLS)	Min (GWR)	1stQu. (GWR)	Median (GWR)	3rdQu. (GWR)	Max (GWR)	Spatial non-stationarity Leung test (sig. level)
Intercept	4.491 ***	1.764	2.392	3.632	4.573	4.971	***
MACM	-0.114 ***	-0.136	-0.120	-0.084	-0.039	-0.013	***
SFW	0.316 ***	0.199	0.294	0.356	0.387	0.452	***
IECM	0.006***	0.005	0.006	0.006	0.007	0.010	n.s.
IMR	-0.006 *	-0.014	-0.008	-0.004	-0.002	0.000	**
MAR	0.041 *	-0.033	-0.010	0.007	0.061	0.089	***
CMR	-0.064	-0.248	-0.145	-0.058	-0.017	0.035	***
$R^2_{OLS} = 0.61$		$R^2_{GWR} = 0.77$		$AIC_{OLS} = -227.28$		$AIC_{cGWR} = -245.50$	
$I_{OLSres} = 0.21 **$		$I_{GWRres} = 0.09$ *n.s.*		$(RSS_{OLS} - RSS_{GWR}) = 0.20 ***(BFC99test)$			
$*** = p < 0.001$		$** = p < 0.01$		$* = p < 0.05$		*n.s. = not significant*	

Datasource : *Geodemo – ISTAT*; *Reference year* : 2005.

References

Brundson C., Fotheringham A. S., Charlton M. (2006). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28, 281-298.

Fotheringham A. S., Charlton M., Brundson C. (2002). *GWR - The analysis of spatially varying relationships*, John Wiley & Sons.

Classification of spatio-temporal series and hidden Markov models

Luigi Spezia

1 Anisotropic and inhomogeneous hidden Markov models for spatio-temporal data on a cylindrical lattice

Hidden Markov models (HMMs) are generalizations of mixture models, obtained by adding a latent, or hidden, Markov process which drives the observed process.

Motivated by a real data problem, an anisotropic and inhomogeneous spatio-temporal HMM with an unknown number of states is made up on a cylindrical lattice. A Bayesian inference procedure, based on a reversible jump Markov chain Monte Carlo algorithm, is proposed to estimate both the dimension and the unknown parameters of the model.

The real data problem is the modelling in time and in space of the concentrations of three dissolved inorganic nitrogens recorded monthly by the Scottish Environmental Protection Agency in the 56 major Scottish rivers. The time series of ammoniacal nitrogen, nitrite, and nitrate are analysed. Samples are collected at the most downstream part of the river which can be conveniently accessed and is above the tidal limit. Further details on the analysis of some time series of dissolved inorganic nitrogens in the Scottish rivers can be found in Spezia et al. (2010).

The 56 gauging stations can be linked to create a circle, and, given that we have a time series for any river, the spatio-temporal data set can be displayed on a cylinder.

The states of the hidden Markov process allows the classification of the observations in a small set of groups. The different states can represent increasing levels of pollution. By means of HMMs, the observations can be clustered without fixing a priori neither the number of groups nor any threshold.

In the Bayesian model presented here, the hidden Markov process is an anisotropic and inhomogeneous Potts model. The Potts model is widely used in statistical mechanics (Chandler, 1987) to model the spins of elementary particles that are placed on a lattice. Here the hidden Potts model is assumed to be anisotropic (i.e., variant under rotations) and inhomogeneous (i.e., variant under translations). Anisotropy is due to the presence of two different parameters describing the link

Luigi Spezia,
Biomathematics & Statistics Scotland, Aberdeen, UK, e-mail: luigi@bioss.ac.uk

between neighbouring pixels: one for the temporal relation and the other for the spatial relation. Inhomogeneity is established by assuming that the spatial relation is a function of the physical distance between two neighbouring sites.

When modelling the concentrations of nitrates and of ammoniacal nitrogen, a periodic component is also added, whereas it is not necessary in the case of nitrites.

Furthermore, there are many missing values in each of the series analysed here. We assume that the missing data are missing at random. Thus the missing data mechanism is ignorable for posterior inference, and this is justified in our application.

The classification is tackled in a fully-Bayesian way because a hierarchical model in the Bayesian framework is made up: the likelihood is used to describe the concentrations of a nitrogen, a Markov prior is introduced to represent suitably the hidden cylindrical lattice, and hyperpriors are elicited on the parameters of the Normal conditional densities of the observations and of the hidden process.

When dealing with Bayesian spatio-temporal HMMs, the highest hurdle to be jumped over is represented by the cumbersome computation of the likelihood, which contains an intractable normalizing constant. Many solutions have been proposed to overcome this problem. In general, the more sophisticated are the approximation procedures, the simpler are the models.

Here, we consider the simplest likelihood approximation available, i.e. the pseudo-likelihood (Besag, 1975), because we want to focus our technical exploration on the segmentation ability of spatio-temporal HMMs, when other sources of complexity are structured: complex neighbourood systems, unknown number of hidden states, continuous interactions, large grids.

To perform our Bayesian inference a RJMCMC algorithm is developed. The parameters are updated by Gibbs sampling and the dimension of the model changes in split-and-merge moves. Our RJMCMC algorithm consists of many sweeps, each one constitued by seven steps: (i) the generation of the means of the Normal densities, (ii) the generation of the precisions of the Normal densities, (iii) the generation of the periodic components, (iv) the generation of the spatial interaction, (v) the generation of the temporal interaction, (vi) the generation of the hidden Markov cylinder, (vii) the split of one component into two or the merge of two components into one.

Current research activities are focussed on the analysis of the univariate spatio-temporal series. Then the multivariate modelling will be considered.

References

Besag J.E. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24, 179-195.

Chandler D.(1987). *Introduction to Modern Statistical Mechanics*, Oxford: Oxford University Press.

Spezia L., Futter M.N., Brewer M.J. (2010). Periodic multivariate Normal hidden Markov models for the analysis of water quality time series. *Environmetrics*, accepted.

Contributed Session 11

Advances in Latent Class Modeling

A latent class version of the inverse probability-to-treatment weighted estimator for dynamic causal effects

Francesco Bartolucci, Fulvia Pennoni and Luca Pieroni

1 Introduction

In many social and economical studies, causal inference on a certain treatment or policy is only possible through observational data. In these contexts, controlling the confounding may be challenging, especially when the treatment varies over time and a reduced set of individual covariates is available. In the latter case, the unmeasured confounding, due to unobservable covariates affecting both the treatment and the response variables, may strongly bias the estimate of the causal effect of this treatment.

For the case of longitudinal data, with time-dependent treatments, an important method to remove the unmeasured confounding is represented by the *inverse probability-of-treatment weighted* (IPTW) estimator (Robins et al., 2000); see also (Gill and Robins, 2001). On the basis of potential outcomes (Rubin, 2005), this method allows us to consistently estimate dynamic causal effects in the presence of time dependent covariates that may be simultaneously confounders. However, as stated in Robins et al. (2000), this method cannot be used to remove the unmeasured confounding. Other relevant approaches are proposed in Lechner and Miquel (2009) and Murphy (2003), which deal with estimators for optimal treatment rules for longitudinal studies. The potential outcome model they propose is based on a series of conditional independence assumptions. Even these methods cannot correct the treatment estimation bias due to the unmeasured confounding.

Francesco Bartolucci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, e-mail: bart@stat.unipg.it

Fulvia Pennoni
Dipartimento di Statistica, Università degli Studi di Milano-Bicocca,
e-mail: fulvia.pennoni@unimib.it

Luca Pieroni
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, e-mail: lpieroni@unipg.it

In this work, we propose an extension of the IPTW estimator, which is able to correct for certain types of unmeasured confounding. This extension is based on the latent regression model and is illustrated in more detail in the following section.

2 Main Results

First of all, we formulate a model on the potential outcomes which may be used in the presence of longitudinal data and accounts for both measured and unmeasured confounding. For a given subject $i = 1, \ldots, n$, let $\mathbf{s} = (s_1, \ldots, s_T)$ denote a sequence of indicator variables for the treatment received at the T occasions of observation. The potential outcomes are denoted by $Y_{it}^{(\mathbf{s})}$ for any possible configuration of $\mathbf{s}$. Then, casual effects of the treatment are expressed as suitable contrasts between $Y_{it}^{(\mathbf{s})}$ and $Y_{it}^{(\mathbf{s}^*)}$, where $\mathbf{s}^*$ is obtained by substituting certain elements of $\mathbf{s}$ by 0. The model is formulated by assuming that both the treatment sequence and these potential outcomes are affected by time-varying covariates collected $\mathbf{X}_{it}$ and by a latent variable U_i representing the effect of unobservable covariates.

In order to estimate the model parameters and then the causal effect of the treatment, we use a method based on two steps:

1. A latent regression model on the observed data is estimated so as to provide a classification of the observed subjects into a reduced number of groups. Subjects in the same group are assumed to have the same behaviour in terms of the effect of the covariates on the treatment choice and on the response variables.
2. For each group of subjects obtained as above, an IPTW estimator is separately applied. The estimated causal effects coming from the different groups are then combined by suitable weighted averages.

The proposed approach is studied by simulation and is illustrated by an application to a dataset concerning particular types of labor market programme.

References

Gill R.D., Robins J.M. (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, **29**, 1785–1811.

Lechner M., Miquel R. (2009). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics* doi:10.100/s00181-009-0297-3.

Murphy S.A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B*, **65**, 331–366.

Robins J.M., Hernn M.A., Brumback B. (2000). Marginal structural models and causal inference. *Epidemiology*, **11**, 550-560.

Rubin D.B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, **100**, 322-331.

Finite mixture modeling of censored longitudinal data

Bettina Grün, Kurt Hornik

1 Introduction

Finite mixture models are a flexible tool for modeling unobserved heterogeneity and finding latent groups in data because different parametric models can be assumed for each component of the mixture to suitably model the data. In a longitudinal setting mixed-effects models are a standard method because they allow to take the repeated observations structure into account and model individual within-group heterogeneity. An additional challenge arises for data including censored observations. Mixed-effects models including censored data have been proposed in Pettitt (1986) and mixture models for binned data in Cadez et al. (2001).

This paper combines finite mixtures of mixed-effects models and mixed-effects models using censored data and proposes the model class of finite mixtures of mixed-effects models for left-censored data. The estimation and inference in a maximum likelihood setting are described by outlining a suitable EM algorithm. The application of the model is illustrated on data of HIV RNA level measurements.

2 Main results

The finite mixture model of mixed-effects models with G components is given for uncensored observations by

$$Y_i \sim \sum_{g=1}^{G} \pi_g N(X_i\alpha_g, Z_i\Psi_g Z_i^\top + \sigma_g^2 I_i). \tag{1}$$

Bettina Grün,
Institute for Statistics and Mathematics, WU Vienna, e-mail: Bettina.Gruen@wu.ac.at

Kurt Hornik,
Institute for Statistics and Mathematics, WU Vienna, e-mail: Kurt.Hornik@wu.ac.at

The component weights π_g are restricted to be positive and sum to one. $N(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and variance-covariance matrix Σ. α_g are the fixed effects parameters with their covariates X_i and Ψ_g the variance-covariance matrix of the random effects with their covariates Z_i. σ_g^2 is the error variance. Due to censoring only (Q_i, C_i) is observed instead of Y_i. Q_i denotes the observed, potentially censored value and C_i the censoring indicator which is zero for uncensored and one for left-censored observations.

Estimation with the EM algorithm relies on the fact that the likelihood of the observed data combined with some latent unobserved variables representing missing information is easier to maximize. In the case of finite mixtures of mixed-effects model and censored data three different kinds of information are missing: (1) the group assignment indicating from which component the individuals are, (2) the random effects and (3) the uncensored observations.

In the E-step the law of the iterated expectation is used by first only integrating out the random effects and then the underlying values of the censored observations and the component memberships. The mean and the variance of the censored observations can be determined as proposed in Tallis (1961) for truncated multivariate normal distributions (see also Vaida and Liu, 2009). Given the expected complete log-likelihood the M-step can be solved in closed form.

The application of the model is illustrated using the HIV-1 viral load data from clinical trial ACTG 315. The mixture model specification allows to investigate whether the same functional relationship holds for all patients or if groups of patients with different HIV RNA level developments exist. Hence, this model can either be used if (1) the a-priori functional relationship is not known or (2) to check if the assumed functional relationship is applicable. Results indicate that the assumption of a monotonically decreasing relationship does not hold for all patients.

References

Cadez I. V., Smyth P., McLachlan G. J., McLaren C. E. (2001) Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1), 7–34.

Laird N. M., Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Pettitt A. N. (1986). Censored observations, repeated measures and mixed effects models: An approach using the EM algorithm and normal errors. *Biometrika*, 73, 635–643.

Tallis G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society B*, 23, 223–229.

Vaida F., Liu L. (2009). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, 18, 797–817.

Determining the number of components in multilevel mixture (factor) models

Roberta Varriale, Jeroen K.Vermunt

1 Introduction

Recently, in social science literature, various types of mixture models have been developed for datasets having a multilevel structure. In two-level datasets, models may include finite mixture distributions at the lower- and/or the higher-level of the analysis. In our project, we investigate the performance of various model selection methods based on some information criteria (IC), such as Bayesian IC (Schwarz, 1978), Aikake's IC (Akaike, 1974), Consistent AIC (Bozdogan, 1987) and AIC3 (Bozdogan, 1993). One difficulty occurring in the use of BIC and CAIC in the context of multilevel models is in choosing the appropriate sample size included in their formula and we investigate whether this should be the number of groups, the number of individuals, or either the number of groups or individuals depending on whether one has to decide about model features concerning the higher or lower level. Starting form the work of Lukociene and Vermunt (2010), we address the issue of determing the correct number of LCs both in multilevel LC models, with categorical latent variables at both levels of the analysis, and in multilevel mixture factor models and latent growth mixture models, with continuous latent variables at the lower-level and a categorical latent variable at the higher-level.

Roberta Varriale,
Department of Statistics "G.Parenti", University of Florence, Viale Morgagni, 59 - 50134 Firenze (Italy) – Department of Methodology and Statistics, Tilburg University, P.O. Box 90153 5000 LE Tilburg (The Netherlands), e-mail: roberta.varriale@ds.unifi.it

Jeroen K.Vermunt,
Department of Methodology and Statistics, Tilburg University, P.O. Box 90153 5000 LE Tilburg (The Netherlands) e-mail: j.k.vermunt@uvt.nl

2 Multilevel LC model, preliminary results

Several types of extensions of the latent class (LC) or mixture models have been developed for the analysis of data sets having a hierarchical structure. The most popular variant is the multilevel LC model with finite mixture distributions at multiple levels of a hierarchical structure (Vermunt, 2003); that is, with LCs for both lower-level units (e.g. individuals, citizens or patients) and higher-level units (e.g. groups, regions, or hospitals). A problem in the application of this model is that determining the number of LCs is much more complicated than in standard LC analysis because it involves multiple, nonindependent, decisions. In our work, we propose a three-step model fitting procedure for deciding about the number of higher- and lower-level classes, as well as investigate the performance of information criteria, such as BIC, AIC, CAIC and AIC3, in the context of multilevel LC analysis, with different types of response variables, namely, categorical and continuous.

The three main conclusions of our simulations studies are that 1) the proposed three-step model fitting strategy works rather well, 2) the number of higher-level units (K) is the preferred sample size for BIC and CAIC, both for decisions about higher- and lower-level classes, and 3) with categorical indicators, AIC3 and BIC based on the higher-level sample size are the preferred measures for deciding about the number of LCs at both the higher and lower level. With continuous indicators, BIC(K) performs better than AIC3. AIC performs best in very specific situations, namely with poorly separated classes and categorical indicators.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

Bozdogan H. (1987). Model selection and Akaike's information criterion(AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

Bozdogan H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In Studies in Classification, Data Analysis,and Knowledge Organization, 40-54, EDS Opitz O., Lausen B., Klar R., Heidelberg: Springer-Verlag.

Lukočienė O., Vermunt J.K. (2010). Determining the number of components in mixture models for hierarchical data. In Advances in data analysis, data handling and business intelligence, 241-249, EDS Fink A., Berthold L., Seidel W., Ultsch A., Berlin-Heidelberg: Springer.

Schwarz G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Vermunt J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Contributed Session 12

Clustering and Dissimilarities

Optimal length choice in top k ordered lists aggregation

Luisa Cutillo, Annamaria Carissimo, Diego di Bernardo

1 Introduction

We consider the problem of combining ranking results from various sources. One of the underlying goals of the rank aggregation methods is to combine many different rank orderings on the same set of candidates in order to obtain a consensus ordering. Rank aggregation has been studied in many disciplines and it finds its origins in the context of the World Wide Web (Dwork et al., 2001). In particular rank aggregation techniques is gaining a growing attention in the bioinformatics applications. As example microarray data can be interpreted as ordered lists of genes and so analyzed via rank aggregation methods as suggested in (DeConde et al., 2006) and (Pihur et al., 2008). Choosing an appropriate measure of dissimilarity between lists, and a reasonable top k length for a particular list at hand remain today the biggest challenges of the aggregating process. Recent available methods are highly performing on lists of limited length but fails when their lengths overcome a certain size. Moreover the choice of the ranked lists length and the top k elements to consider in the aggregation process is left as an exogenous variable. We introduce a method to automatically select the ranked lists to consider together with the top-k elements that would result in the best informative aggregate list. We will show an application to biological microarray data.

Luisa Cutillo,
Telethon Institute of Genetics and Medicine, Napoli and University of Naples Parthenope,
e-mail: cutillo@tigem.it

Annamaria Carissimo,
Telethon Institute of Genetics and Medicine, Napoli, e-mail: carissimo@tigem.it

Diego di Bernardo
Telethon Institute of Genetics and Medicine, Napoli and University of Naples Federico II,
e-mail: dibernardo@tigem.it

2 Main results

There are several standard methods to define a distance measure between two lists. We choose the Kendall's tau metric. Let L_i indicate a generic ordered list an let:

$$r^{L_i} : a \in L_i \longrightarrow r^{L_i}(a) \in \{1 \dots |L_i|\} \tag{1}$$

be the ranking function of the list L_i. Let $L_i(h)$ be the top h sublist of L_i. Suppose we have N ordered lists $L_i\{i = 1, \dots, N\}$ whose lengths, $k_i = |L_i|\{i = 1, \dots, L\}$, are not necessary the same. Our method iteratively works on pairs of full or partial ranked lists which are clustered according to the a modified version of the Kendall's tau distance (Pihur et al., 2008):

$$K(L_i, L_j) = \sum_{t,u \in L_i \cup L_j} K^p_{tu} \tag{2}$$

where $K^p_{tu} : L_i x L_j - > \{0, 1, p\}$ is a piece-wice function of the relative orderings (1).

Our choice of $p = \frac{|L_i \cup L_j| \setminus |L_i \cap L_j|}{|L_i| \cup |L_j|}$ accounts for the relative mismatches of the two lists. We hereby describe our iterative algorithm to create a consensus list in each cluster. The core of each step is the definition of a pair of lengths $(\hat{k}_i, \hat{k}_j)$ s.t. the aggregation of the two top k lists $L_i(\hat{k}_i)$ and $L_j(\hat{k}_j)$ results in the best consensus list of the pair considered. This process consists in exploring and scoring a finite set of possible $(\hat{k}_i, \hat{k}_j)$ values through a bayesian test based on the hypergeometric distribution. The aggregation step is then performed via standard literature aggregation methods for partial lists.

The overall method is applied to a public dataset, the *Connettivity Map* (cMap), consisting of 6000 lists of 22000 genes ranked according to their differential expression with respect to 1309 different treatments on five different human cell lines (Lamb et al., 2006).

References

DeConde R., Hawley S., Falcon S., Clegg N., Knudsen B., Etzioni R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol*, 5,Article 15.

Dwork C., Kumar R., Naor M., Sivakumar D. (2001). Rank aggregation methods for the Web. *Proceedings of the 10th international conference on World Wide Web*

Lamb J., Crawford E.D., Peck D., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* (313),1929–1935.

Pihur V., Datta S., Datta S. (2008). Finding cancer genes through meta-analysis of microarray experiments: Rank aggregation via the cross entropy algorithm. *Genomics*, (92),400–403.

Recognition of ultrametrics and tree-metrics in optimal time

Bernard Fichet

1 Introduction

Recall that a dissimilarity d on an n-set X is said to be ultrametric if it obeys the *ultrametric inequality*:

$$d(x,y) \leq max[d(x,z),d(y,z)] \tag{1}$$

for every x,y,z in X. Equivalently, one may suppose that x,y,z are distinct, and in that case, the definition may be extended to *pre-dissimilarities* and *pre-ultrametrics*, that is functions d with possible negative values. It is well-known that ultrametrics are in one-to-one correspondence with indexed hierarchies, so producing a nice visual display through a rooted tree. Many algorithms in $O(n^2 logn)$ for the recognition of ultrametricity have been developed. More recently, Heun (2008), has proposed an optimal algorithm, with complexity $O(n^2)$. His algorithm is recursive and builds the associated rooted tree. We here propose another, and very simple, optimal algorithm, based on the search of a *compatible order*, to recognise whether or not a dissimilarity is ultrametric, and if so to build the hierarchy.

Metrics *of tree-type*, that is metrics d such that (X,d) is isometrically embeddable in a weighted tree, generalise ultrametrics. The following *four-point condition* characterises metrics of tree-type (Buneman, 1974):

$$d(x,y)+d(z,t) \leq max[d(x,z)+d(y,t),d(x,t)+d(y,z)] \tag{2}$$

for every x,y,z,t in X. Actually, there is a strong link between ultrametricity and tree-metricity, via the so-called *Farris transformation* . It requires the choice of a base-point a in X, and up to an additive constant, it is defined by:

$$d^a(x,y) = d(x,y) - d(a,x) - d(a,y) \tag{3}$$

Bernard Fichet,
LIF Marseilles University, e-mail: Bernard.Fichet@lif.univ-mrs.fr

on every pair of distinct elements x, y in X. Note, that d^a may be negative. Then, it has been established that a metric d is of tree-type if and only if d^a is pre-ultrametric, see Bandelt (1990) and many others.

2 Main results

Our algorithm for ultrametricity derives from a very simple property. Given d, we define:

- A linear order on X:
  select any element, say x_1, in X. Having the sequence $x_1 < x_2 < \cdots < x_k, (k < n)$, and noting $X_k := X \setminus \{x_1, \cdots, x_k\}$:
  let x_{k+1} in X_k realise $d(x_k, x_{k+1}) = min\{d(x_k, y) : y \in X_k\}$
- A dissimilarity d_* on X:
  put $a_i := d(x_i, x_{i+1}), i = 1, \cdots, n-1$. Define d_* by:
  $\forall i, j = 1, \cdots, n, i < j, d_*(x_i, x_j) = max\{a_k : k = i, \cdots, j-1\}$

  Then the following property is proved.

Property 1. d is ultrametric if and only if $d = d_*$.
If d is ultrametric, the order is compatible with d.

Clearly, the chain $x_1, \cdots, x_n$ is built in $O(n^2)$ time. A similar complexity holds for computing d_*. It suffices to observe that for $i < j < j+1, d_*(x_i, x_{j+1}) = max[d_*(x_i, x_j), d_*(x_{i+1}, x_{j+1})]$.

Still, the indexed hierarchy, hence the rooted tree, is built in $O(n^2)$ time, since the clusters are some intervals of the compatible order.

Using (3), it is easy to check in $O(n^2)$ time whether a metric d is of tree-type. That seems much more complicated if metricity of d is not assumed. However, that may be done with the same complexity, if we apply a local metric condition, as established in Chepoi and Fichet (2000).

References

Bandelt H.J. (1990). Recognition of tree metrics. *SIAM Journal of Discrete Mathematics*, 3, 1-6.

Buneman P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory, Ser. B*, 17, 48-50.

Chepoi V., Fichet B. (2000). l_∞-Approximation via Subdominants. *Journal of Mathematical Psychology*, 44, 600-616.

Heun V. (2008). Analysis of a modification of Gusfield's recursive algorithm for reconstructing ultrametrics trees. *Information Processing Letters*, 108, 222-225.

A dissimilarity measure between two hierarchical clusterings

Isabella Morlini, Sergio Zani

1 Introduction

This paper deals with the comparison of two hierarchical clusterings obtained from the same set of n objects, using different linkages, different distances or different variables. In the literature, the most popular measures have been proposed for comparing two partitions at a certain stage of two hierarchical procedures (Rand, 1971; Fowlkes and Mallows, 1983; Hubert and Arabie, 1985). We propose a new index for measuring the global dissimilarity between two hierarchical clusterings. This index can be decomposed into the contributions pertaining to each stage of the hierarchies.

2 The index and its properties

Suppose we have two hierarchical clusterings of the same number of objects, n. Let consider the $N = n(n-1)/2$ pairs of objects and let define, for each non trivial partition in k groups ($k = 2,\ldots,n-1$), a binary variable X_k with values $x_{ik} = 1$ if objects in pair i ($i = 1,\ldots,N$) are classified in the same cluster in partition in k groups and $x_{ik} = 0$ otherwise. A binary $(N \times (n-2))$ matrix $\mathbf{X}_g$ for each clustering g ($g = 1,2$) may be derived, in which the columns are the variables X_k. A global measure of dissimilarity between the two clusterings may be defined as follows:

$$Z = \frac{\| \mathbf{X}_1 - \mathbf{X}_2 \|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|} \tag{1}$$

Isabella Morlini,
Dipartimento di Economia, Università di Modena e Reggio Emilia,
e-mail: isabella.morlini@unimore.it

Sergio Zani,
Dipartimento di Economia, Università di Parma. e-mail: sergio.zani@unipr.it

where $\| \mathbf{A} \| = \sum_i \sum_k \| a_{ik} \|$ is the L_1 norm of the matrix $\mathbf{A}$. In this case, since the values in the matrix are binary, the L_1 norm is equal to the square of the L_2 norm.

Index Z has the following properties.

- It is bounded in [0,1].
- $Z = 0$ if and only if the two hierarchical clusterings are identical and $Z = 1$ when the two clusterings have the maximum degree of dissimilarity, that is when for each partition in k groups and for each i, objects in pair i are in the same group in clustering 1 and in two different groups in clustering 2 (or vice versa).
- It is a distance, since it satisfies the conditions of non negativity, identity, symmetry and triangular inequality.
- The complement to 1 of Z is a similarity measure, since it satisfies the conditions of non negativity, normalization and symmetry.
- It does not depend on the group labels since it refers to pairs of objects.
- It may be decomposed in $(n-2)$ parts related to each pair of partitions in k groups since:

$$Z = \sum_k Z_k = \sum_k \sum_i \frac{|x_{1ik} - x_{2ik}|}{\| \mathbf{X}_1 \| + \| \mathbf{X}_2 \|} \tag{2}$$

  The plot of Z_k versus k shows the distance between the two clusterings at each stage of the procedure.
- The numerator of Z_k in (2) can be expressed as a function of the quantities in Table (1) and can be related to the Rand index, R_k (see Warrens, 2008, for the formula of R_k in terms of the quantities in Table (1)):

$$\sum_{i=1}^{N} |x_{1ik} - x_{2ik}| = P_k + Q_k - 2T_k = N(R_k - 1) \tag{3}$$

Table 1 *Contingency table of the cluster membership of the N object pairs*

	Second clustering (g=2)		
First clustering (g=1)	Pairs in the same cluster	Pairs in different clusters	Sum
Pairs in the same cluster	T_k	$P_k - T_k$	P_k
Pairs in different clusters	$Q_k - T_k$	$U_k = N - T_k - P_k - Q_k + 2T_k$	$N - P_k$
Sum	Q_k	$N - Q_k$	$N = n(n-1)/2$

References

Fowlkes E. B., Mallows C. L. (1983). A method for comparing two hierarchical clusterings. *JASA*, 78, 553–569.

Hubert L. J., Arabie P., (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Rand W. M. (1971). Objective criteria for the evaluation of clustering methods. *JASA*, 66, 846–850.

Warrens M. J. (2008). On the equivalence of Cohen's Kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, 25, 177–183.

Contributed Session 13

Customer Satisfaction and Conjoint Analysis

Unveiling non-linear relationships between perceived satisfaction and quality

Giuseppe Boari, Gabriele Cantaluppi

1 Introduction

Structural equation models with latent variables consider the following relations

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

$$\mathbf{x} = \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\xi} + \boldsymbol{\delta} \qquad \mathbf{y} = \boldsymbol{\Lambda}_{\mathbf{y}}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

where the inner model (1) states the structural linear relationship among exogenous latent variables, $\boldsymbol{\xi}$, and the endogenous ones, $\boldsymbol{\eta}$, explained by the matrices of coefficients $\mathbf{B}$, lower triangular, and $\boldsymbol{\Gamma}$; the outer measurement model (2) defines the linear reflective relationships among the latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, and the corresponding manifest variables $\mathbf{x}$ and $\mathbf{y}$. All previous variables are defined to be the differences from their average values. In several applications the previous relations do not appear to be of the linear type; here we will deal with unveiling non-linearities in the measurement model (2); see Boari et al. (2008) and Fuller et al. (2006) for a wider analysis. For example, non-linear relationship are supposed to exist between perceived satisfaction and perceived quality, with reference to the facilities of the branch offices of a financial service network. Following Kano et al. (1996) the non-linearity may be distinguished between "basic type" and "attractive type", whose graphical representation is given in Fig. 2.

In Boari et al. (2008) two different procedures are used to identify the presence of non-linearity. The result showed that the attribute "neatness of the offices" may be classified as "basic", leaving customers not particularly satisfied when fulfilled, but dissatisfied if not. As alternative approach we propose to perform the rank transformation of the proxy variables suspected to have a non-linear link with the corresponding latent ones, and to identify the actual nature of the relationship.

Giuseppe Boari, Gabriele Cantaluppi
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano,
e-mail: {giuseppe.boari, gabriele.cantaluppi}@unicatt.it

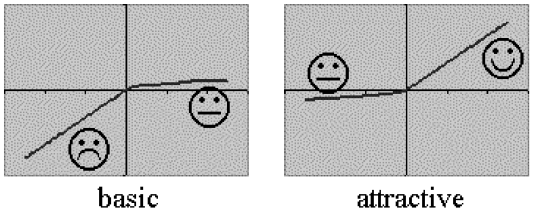


Fig. 1 Types of Kano non-linear relationships left: must-be or expected right: exciting

2 The Main Proposal

Let us assume, as usual, model (2) be of factor complexity one, that is each manifest variable depends exclusively on only one latent variable. So we may focus our attention on a single latent construct, say ξ, in order to unveil if a non-linear relationship exists with some of its indicators, say $x_1, \ldots, x_k$. Should the latent scores $\hat{\xi}$ be available (possibly obtained by means of the PLS algorithm or by calculating the first principal component), one may compute the correlation coefficients between those scores and each observed manifest variable or, more specifically, consider the reliability procedure: this allows one to obtain the Cronbach's α coefficient and its modifications (α-if) when an item is dropped, making it possible the so-called scale purification. It is well-known that when the relationship between $\hat{\xi}$ and, say, x_i is a non-linear one, the correlation between the rank transformed variables $\text{rank}(\hat{\xi})$ and $\text{rank}(x_i)$ is greater than the correlation between the original variables $\hat{\xi}$ and x_i. Should the previous result occur for x_i, a suspect of a non-linear relation will stem. This may be confirmed by comparing the reliability results obtained for both the original manifest variables and the ranked ones: a dissimilarity in the variables to be dropped may be interpreted as an indication of non-linearity. The procedure may be generalized to the global model (1), (2), by estimating the latent scores by means of the PLS algorithm, first with reference to the original data, and then to the rank transformed variables. Observe that the α-if criterion may be correctly applied, being Cronbach's α a lower bound reliability index for congeneric measures.

References

Boari G., Cantaluppi G., Bertelli S. (2008). Non-linear relationships in SEM with latent variables: some theoretical remarks and a case study. In: First joint meeting of the Société Francophone de Classification and CLADAG, Book of Short Papers, Napoli: Edizioni Scientifiche Italiane, 201–204.

Füller J., Matzler K., Faullant R. (2006). Asymmetric Effects in Customer Satisfaction. *Annals of Tourism Research*, 33 (4), 1159–1163.

Kano N., Seraku N., Takahashi F., Tsuji S. (1996). Attractive Quality and Must-Be Quality. In: Hromi, J.D. EDS The Best on Quality. International Academy for Quality, Vol. 7, ASQ Quality Press, Milwaukee, 165–186.

Ordinal logistic regression for the estimate of the response functions in the conjoint analysis

Amedeo De Luca

1 Introduction

The Conjoint Analysis (COA) model proposed here - an extension of the traditional COA approach - is based on overall desirability categories k $(k = 1, 2, ..., K)$ chosen from a sample of respondendts, for each of S hypothetical product profiles. The ordinal variable response Y_k (i.e. evaluation of the overall) is described by an ordered logit model, that directly incorporates the order of the categories of the Y_k.

The main characteristic distinguishing the proposed model from the traditional full-profile COA (Green, 1971) is that in the first one the respondent expresses preferences by choosing the overall among K desirability categories, rather than by rating or ranking distinct product profiles, in the second one. The approach also differs from the "Choice-Based Conjoint" Analysis (CBC) model in which the respondent expresses preferences by choosing concepts from sets of concepts.

To link the categories of overall evaluation to the factor levels, we adopt a cumulative logit model (De Luca, 2006) at the aggregate level (*pooled model*) (Moore, 1980).

The novelty value in our approach is that one set of aggregated part-worths is estimated in connection with each overall category Y_k, as many as the overall ordered categories are (K), unlike the traditional metric and non metric COA and CBC analysis, which give only one response function. Moreover, the proposed model provides the following advantages: the use of the probability P_{ks} as an average response, which does not require preliminary scale adjustments to render the preference scale "metric" (in the non metric COA), and a cross-check of the effects of the attribute levels on the different k categories.

Amedeo De Luca,
Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore di Milano,
e-mail: amedeo.deluca@unicatt.it

2 Main results

The number of profiles S, resulting from the total number of possible combinations of levels of the M attributes or factors (X) of a product, constitute a $full-factorial$ experimental design. It is assumed the *overall* evaluation (Y) of a product consists in the choice of one of the ordered categories $k = 1,2,...,K$.

The effects of the factors express the variations of the probabilities P_{ks} associated with the vector $\mathbf{z}_s$ corresponding to the combination s $(s = 1,2,...,S)$:

$$P_{ks}(Y_k = 1|\mathbf{z}_s) = \pi_k(\mathbf{z}_s) = \exp(\underline{\delta}'_k\mathbf{z}_s)/[1+\exp(\underline{\delta}'_k\mathbf{z}_s)] \quad (1)$$

where $\mathbf{z}_s$ is the vector of the dummy explanatory variables relative to the combination or profile s; $\underline{\delta}'_k$ is the unknown vector of regression coefficients of the factors.

The k-th cumulative response probability is:

$$P_{ks}(Y_k \leq k|\mathbf{z}_s) = F_k(\mathbf{z}_s) = \pi_1(\mathbf{z}_s) + \pi_2(\mathbf{z}_s) + ... + \pi_k(\mathbf{z}_s) \quad (2)$$

In the model the K-th equation can be drawn from the remaining $q = K-1$ equations.

The cumulative logits of the first $(K-1)$ cumulative probabilities are:

$$L_k(\widetilde{\mathbf{z}}_s) = \ln\left[\frac{F_k(\widetilde{\mathbf{z}}_s)}{1-F_k(\widetilde{\mathbf{z}}_s)}\right] = \ln\left[\frac{\pi_1(\widetilde{\mathbf{z}}_s)+\pi_2(\widetilde{\mathbf{z}}_s)+...+\pi_k(\widetilde{\mathbf{z}}_s)}{\pi_{k+1}(\widetilde{\mathbf{z}}_s)+\pi_{k+2}(\widetilde{\mathbf{z}}_s)+..+\pi_K(\widetilde{\mathbf{z}}_s)}\right] = \delta_k + \widetilde{\underline{\delta}}'\widetilde{\mathbf{z}} \quad (3)$$

with $k = 1,2,...,q$; $\widetilde{\mathbf{z}}_s$ is the vector of the reduced matrix $\widetilde{\mathbf{Z}}$ (in which it has been dropped out one of the dummy variables for each level of the M factors X); δ_k is the intercept parameter associated to the reference category; $\widetilde{\underline{\delta}}'$ is the vector of the unknown coefficients. In equation (3) $\widetilde{\underline{\delta}}'$ does not have a k subscript (*Proportional Odds Assumption-POA*), so the model assumes the same effects as $\widetilde{\mathbf{Z}}$ for all $K-1$ on all cumulative logit results, in a parsimonious model for ordinal data.

We provide an application to real data with PLUM-Ordinal regression procedure of SPSS and an interpretation of the main effects of the model.

References

De Luca A. (2006). A Logit Model with a Variable Response and Predictors on an Ordinal Scale to Measure Customer Satisfaction. *Quality and Reliability Engineering International*, 22, 591-602.

Green P.E., Rao V.R. (1971). Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research*, 8, 355-363. 591-602.

Moore W.L. (1980). Levels of Aggregation in Conjoint Analysis: an Empirical Comparison. *Journal of Marketing Research*,17, 516-523.

Individual Self Balancing Conjoint (ISBC): an adaptive design technique for choice based conjoint

Peter Kurz, Andrzej Sikorski

1 Introduction

Current research often addresses how choice tasks should be constructed, because the choice design efficiency has an effect on the validity of the estimates. Choice designs must fulfil four criteria: level balance, orthogonality, minimal overlap and utility balance. The first three can be addressed and tested in a static environment. Utility balance requires prior assumptions or pretests about the underling utility distribution. Recent research shows more and more dynamic approaches that do not rely on prior knowledge and consider respondents heterogeneity by adapting choice designs individually during the course of the questionnaire. Two major research streams produce different results: Johnson, Huber and Orme do not detect significant improvement with their approach, whereas Toubia, Hauser and Simester demonstrate the feasibility of their FastPACE approach. For the findings however, not many empirical studies confirm the results.

2 Main results

This paper tackles a method developed by the authors to implement the ideas of Toubia, Hauser and Simester into the framework of a surveying software. The structure of the paper is the following: After detailing the algorithms used for the adaptive approach it outlines the research design of the empirical studies, followed by an interpretation of the results. Finally, the paper concludes some limitations and implications of this approach.

Kurz Peter,
TNS Infratest Models and Methods, e-mail: peter.kurz@tns-infratest.com

Sikorski Andrzej,
Technical University Poznan e-mail: andrzejs@et.put.poznan.pl

The ISBC method follows the lines of Toubia et al. (polyhedral method) and this work is used as a reference for the concepts not defined herein. The general idea is that respondents answer the survey questions, providing information about their utility values. These partial results are processed in the background by a specialized version of a maximum likelihood estimator (MLE) that dynamically adjusts current estimates and enforces feasibility constraints (Toubia, 2003). The most significant differences with respect to the original method are the following:
The utilities are estimated with an extended MLE method that we use to generate utilities and their covariance. The MLE step for ISBC s has been optimized with respect to its performance. The feasibility constraints are enforced by means of our algorithm that generates vectors of constrained, normally distributed variables in O(k) time, where k represents vector length. This particular feature gives our method a performance boost. The eigenvectors of the covariance matrix (only directions thereof) are used as ellipsoid axes for choice task generation (Toubia, 2003).

The first empirical study was conducted in autumn 2008 and compares the ISBC approach with ACBC and CBC from Sawtooth Software. The study deals with eco labeling in the automotive industry. The subject is described by 11 attributes with a total of 45 levels. The research design is rather complex for a full-profile exercise but the information in the subsequent choice sets was realistic for such markets. The study contains three experimental sets to test the three methods. For each condition a representative sample of 300 respondents is drawn. The next wave of empirical studies is based on five car clinics from a large German manufacturer. The ISBC exercise contains 12 attributes with up to 12 levels each. The respondents have to answer 20 choice sets in average.

The main difference between the adaptive and the static designs is that the respondents dont get so much arbitrary choice sets. In static designs many of the choice sets could be answered itself or there are no interesting products at all in the choice set. Especially in the complex choice sets of the car clinic we could see, that the involvement of the respondents is much higher when they see, the algorithm is learning their interests and generate more and more choice set which comes close to their individual preferences. For practitioner its not the higher statistical efficiency of the experiment - its the fact, that the validity of respondent answers is much higher and the noise in the data could be reduced significantly that makes this approach interesting.

References

Toubia O. et al. (2003). Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis. *MIT Working Paper 4385-03*.

Contributed Session 14

Recent Trends in Classification

Clustering highly evolving data streams

Antonio Balzanella, Rosanna Verde, Yves Lechevallier

1 Introduction

In recent years a wide number of domains is generating temporally ordered, fast changing, potentially unbounded data streams. Some examples are web data, network traffic, electricity consumption, remote sensor data. Traditional data mining methods fail at dealing with such data because they require the availability of the whole data set and the possibility to perform multiple scans. To overcome these issues, proper strategies have to be developed which update, in an incremental and on-line way, the knowledge about data (a wide review of the challenges in data stream analysis is available in Gama and Gaber (2007)).

In this framework we deal with the challenge of data stream clustering Balzanella et al. (2010), introducing a new strategy in order to discover homogeneous groups of streams and to provide the clustering structure over user specified time horizons. A set of profiles is also achieved to summarize the data behaviors over the time.

Let $S = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ be a set of n streams

$$Y_i = [(y_1, t_1), \ldots, (y_j, t_j), \ldots, (y_\infty, t_\infty)]$$

made by real valued, temporally ordered observations on a discrete time grid

$$T = \{t_1, ..., t_j, ...t_\infty\} \in \Re.$$

A time window w_f with $f = 1, \ldots, \infty$ is an ordered subset of T having size w_s. Each time window w_f frames a subset Y_i^w of Y_i called subsequence where $Y_i^w =$

Antonio Balzanella,
Second University of Naples, e-mail: antonio.balzanella@gmail.com

Rosanna Verde,
DEM-Second University of Naples, e-mail: rosanna.verde@unina2.it

Yves Lechevallier,
INRIA Rocquencourte-mail: yves.lechevallier@inria.fr

$\{y_j, \ldots, y_{j+w_s}\}$. The objective is to find a partition P of S into C clusters such that each stream Y_i belongs to a cluster C_k with $k = 1, ..., C$ and $\bigcap_{k=1}^{C} C_k = \phi$.

The incoming parallel streams are, at first, split into non overlapping windows of fixed size. On the subsequences Y_i^w of Y_i framed by each window w_f a Dynamic Clustering Algorithm (DCA) extended to complex data De Carvalho et al. (2004) is run, in order to get a local partitioning $P_w = C_1^w \cup \ldots \cup C_\kappa^w \cup \ldots \cup C_K^w$ into K clusters and the associated set of prototypes $B^w = (b_1^w, \ldots, b_\kappa^w, \ldots, b_K^w)$ which summarize the behaviors of the streams in time localized windows.

For each local partitions P_w we update a proximity squared matrix $\boldsymbol{A}_{nxn}$ whose items $a_{i,l} = a_{l,i} \in \Re$ (with $i, l = 1, \ldots, n$) maintain the similarity values among the streams.

The main idea is to store in each cell $a_{i,l}$ the number of times each couple of streams is allocated to the same cluster of a local partition P_w. For instance, let us assume to have five streams $(Y_1, Y_2, \ldots, Y_5)$ and a local partition $P_1 = (Y_1^w, Y_2^w)(Y_3^w, Y_4^w, Y_5^w)$, the updating of A consists in adding the value 1 to the cells $(a_{1,2})$, $(a_{2,1})$, $(a_{3,4})$, $(a_{4,3})$,$(a_{3,5})$,$(a_{5,3})$, $(a_{4,5})$ and $(a_{5,4})$

When this procedure is performed on a wide number of windows, it provides an incrementally computed measure of consensus between the couples of streams.

On user demand or at predefined time stamps, the on-line updating of the proximities among streams is stopped in order to get a final partitioning of the streams.

This is obtained, representing the streams as points in a lower dimensional space by means of a Non Metric MultiDimensional Scaling (MDS) on the proximity matrix A and then running a standard k-means algorithm on factorial coordinates of the obtained data points.

It is possible to note that the proximity matrix A can be seen as the adjacency matrix of a weighted graph. According to Bavaud (2006) this involves that running the MDS on A provides very similar results to the Spectral clustering algorithm which is often used for partitioning weighted graphs into minimally interacting components.

References

Balzanella A., Lechevallier Y., Verde R. (2010). Clustering multiple data streams. In: New Perspectives in Statistical Modeling and Data Analysis (S. Ingrassia, R. Rocci, M. Vichi Eds.), Springer.

Bavaud, F. (2006). Spectral clustering and multidimensional scaling: a unified view. In: Data science and classification, Springer. 131–139

De Carvalho F., Lechevallier Y., Verde R. (2004). Clustering methods in symbolic data analysis. In: Banks D, House L, McMorris FR, Arabie P, Gaul E (eds) Classification, clustering, and data mining applications. Springer:Berlin.

Gama J., Gaber, M.M. EDS (2007). Learning from Data Streams: Processing Techniques in Sensor Networks. Ed. Springer Verlag.

Extending SOM with efficient estimation of the number of clusters

Guénaël Cabanes and Younès Bennani

1 Introduction

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. A clustering problem can be defined as the task of partitioning a set of objects into a collection of mutually disjoint subsets. Patterns in the same cluster should be similar to each other, while patterns in different clusters should not (internal homogeneity and the external separation).

An efficient method is the Self Organizing Map (SOM Kohonen, 2001). The SOM is a neuro-computational algorithm to map high-dimensional data to a two-dimensional space through a competitive and unsupervised learning process. The representation of different number of groups in this approach is usually detected using others clustering techniques such K-means or hierarchical methods. In the first phase of the process, the standard SOM approach is used to compute a set of reference vectors (prototypes) representing local means of the data. In the second phase, the obtained prototypes are used to form the final partitioning of data using a traditional clustering method. Such approach is called a two-level clustering method.

One of the most crucial questions in many real-world cluster applications is how to determine a suitable number of clusters K, also known as the model selection problem. Without a priori knowledges there is no simple way of knowing this number. The purpose of this work is to provide a simultaneous two-level clustering approach using SOM, by learning at the same time the structure of the data and its segmentation, using both distance, density and connectivity informations.

Guénaël Cabanes and Younès Bennani,
LIPN-CNRS, UMR 7030, University of Paris 13, e-mail: cabanes@lipn.univ-paris13.fr

2 Main results

This new clustering algorithm assumes that a cluster is a dense region of objects surrounded by a region of low density (Ester et al., 1996; Ultsch, 2005; Pamudurthy et al., 2007). This approach is very effective when the clusters are irregular or intertwined, and when noise and outliers are present. The proposed clustering algorithm divides automatically a given dataset into a collection of subset (clusters), i.e., the number of clusters is determined automatically during the learning process, i.e., no a priori hypothesis for the number of clusters is required.

The dataset is modeled using a SOM with prototypes enriched using knowledges abstracted from the dataset. The idea is first to construct an abstract representation of the data, which is supposed to capture the essential structure of the dataset, then to use this abstraction to find automatically the segmentation. A great advantage of the proposed algorithm, compared to the common partitional clustering methods, is that it is not restricted to convex clusters but can recognize arbitrarily shaped clusters and touching clusters. The validity and the stability of this algorithm are superior to standard clustering methods. This is demonstrated on a set of critical clustering problems and shows excellent results compared to usual approaches. Some extensions have been done to deal with interval data, structured data (kernel method), bi-clustering and so on ...

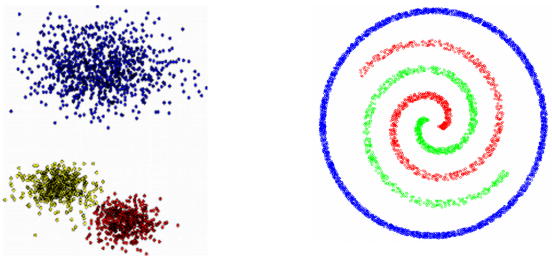


Fig. 1 Example of automatic clustering obtained with DS2L-SOM

References

Ester M., Kriegel H.P., Sander J., Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, 226-231.

Kohonen T. (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin.

Pamudurthy S.R., Chandrakala S., Sakhar C.C. (2007). Local density estimation based clustering. In: Prodeeding of International Joint Conference on Neural Networks, 1338-1343.

Ultsch A. (2005). Clustering with SOM: U*C. In: Procedings of the Workshop on Self-Organizing Maps, 75-82.

A classification approach for structure discovery in search spaces of combinatorial optimization problems

Daniel Cosmin Porumbel, Jin-Kao Hao, Pascale Kuntz

1 Introduction

To overcome the difficulties of developing efficient approaches for large-scale optimization problems, an emergent research area showing a growing interest consists in combining data mining and machine learning with optimization (Boyan et al., 2000; Battiti et al., 2008). Indeed, for large instances of combinatorial optimization problems where the huge size of the search spaces prevents any efficient enumeration, classification of some samples can provide efficient insight into the problem structure. It is well-known that competitive heuristic algorithms take into consideration, either explicitly or implicitly, the structure of the search space, and so, the additional information provided by data analysis can guide the heuristic strategies more efficiently.

This communication illustrates the interest of such an approach in the framework of the k-coloring problem. Let us recall that a k-coloring instance is defined by a graph $G(V,E)$ and a number of colors k. The search space contains all possible vertex coloring with k colors—a coloring is a *partition* of V into k disjoint subsets (classes). The objective is to find a coloring minimizing the number of edges with both ends in the same class (of the same color). Algorithms based on the Tabu Search meta-heuristic represent an effective approach for this problem (Galinier and Hertz, 2006). However, for some instances, even quite sophisticated Tabu search processes can be blocked in local optima. To improve the heuristic design, we have performed a "cartography" of the search space so as to investigate the spatial distribution of

Daniel Cosmin Porumbel,
LERIA, Université d'Angers, 2 Bd Lavoisier, Angers, France
e-mail: porumbel@info.univ-angers.fr

Jin-Kao Hao,
LERIA, Université d'Angers, 2 Bd Lavoisier, Angers, France e-mail: hao@info.univ-angers.fr

Pascale Kuntz,
LINA, Polytech'Nantes, rue Christian Pauc, Nantes, France e-mail: pascale.kuntz@univ-nantes.fr

the potential solutions and their clustering. Here, the data consists of a sample of highest quality potential solutions, defined as partitions of V, reached by the Tabu Search algorithm in a certain time window.

2 Main results

As usually in classification, a dissimilarity function on the data set is required. However, here, the data represents combinatorial objects and the challenge is to introduce an appropriate distance function with a low computing complexity. For the graph coloring problem, the transfer distance between partitions is interesting (Charon et al., 2006), but it is generally calculated using the Hungarian algorithm, resulting in a complexity of at least $O(k^2 + |V|)$. To apply it to search space analysis, we have recently proposed a Las Vegas algorithm that reduces the complexity to $O(|V|)$ (the same as for a Hamming distance) if certain conditions are satisfied. Using this distance and a classical Multidimensional Scaling approach, we have experimentally shown that, for numerous instances, the highest quality solutions are not randomly scattered in the search space, but rather grouped in clusters within spheres of specific diameter.

Using this information, we have introduced in Tabu Search a component which records the heuristic trajectory (by memorizing the visited clusters/spheres), and guides it towards certain high-quality spheres. We have experimentally shown on the DIMACS instances—which are systematically used as test cases for the k-coloring problem—that our approach reaches very competitive results compared to previous local search. The proposed algorithm can even compete with more complex population-based algorithms and improve some of the best-known solutions from the literature (Porumbel et al., 2010).

References

Battiti, R., R. Brunato, and F. Mascia, 2008: *Reactive Search and Intelligent Optimization*. Springer.

Boyan, J., W. Buntine, and A. Jagota, 2000: Statistical machine learning for large-scale optimization. *Neural Computing Surveys*, **3**(1), 1–58.

Charon, I., L. Denoeud, A. Guénoche, and O. Hudry, 2006: Maximum transfer distance between partitions. *Journal of Classification*, **23**(1), 103–121.

Galinier, P. and A. Hertz, 2006: A survey of local search methods for graph coloring. *Computers and operations research*, **33**(9), 2547–2562.

Porumbel, C., J. Hao, and P. Kuntz, 2010: An evolutionary approach with diversity guarantee and well-informed grouping recombination for graph coloring. *Computers and Operations Research*, **37**, 1822–1832.

Contributed Session 15

Regression and Factor Analysis

Generalized extreme value regression in rare events

Raffaella Calabrese, Silvia Angela Osmetti

1 Introduction

We aim at proposing a Generalized Linear Model (GLM) whose link function defined by the Generalized Extreme Value (GEV) distribution with binary dependent variable Y. We define this model as GEV regression. The goal of this paper is to overcome the drawbacks shown by the logistic regression in rare events: the probability of rare events is underestimated and the logit link is a symmetric function, so the response curve approaches 0 as the same rate it approaches 1. If the dependent variable represents a rare event or an extreme event, a symmetric link function is not appropriated. In the extreme value theory the GEV distribution is used to model the tail of a distribution ((Kotz et al., 2000)). Since we focus our attention on the tail of the response curve for the values close to 1, we chose the GEV distribution. The Gumbel and Weibull distributions represent particular cases of the GEV distribution. In GLM (Agresti, 2002) the log-log link function, related to the Gumbel distribution, is used since it is asymmetric. We describe the methodology to estimate the parameters of a GLM model by using a log-log link function related to the Gumbel distribution. By applying a similar procedure, we propose the Weibull regression whose link function is defined by the Weibull distribution. Finally, we generalized the previous models by the GEV distribution, so we define the GEV regression. After computing the likelihood and the score functions, we estimate the parameters by the maximum likelihood method using a Newton algorithm. In order to compute the initial point estimates of the parameters we consider the results obtained for the Gumbel and the Weibull models. Finally, we apply the GEV regression to empirical data on Italian firms to model the default probability.

Raffaella Calabrese,
Department of Quantitative Methods, University of Milan-Bicocca,
e-mail: raffaella.calabrese1@unimib.it

Silvia Angela Osmetti,
Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore,
e-mail: silvia.osmetti@unicatt.it

2 Main results

The GEV cumulative distribution function is given by

$$F_X(x) = \exp\left\{-\left[1-\tau(x)\right]^{-\frac{1}{\tau}}\right\} \quad -\infty < \tau < \infty; \quad 1+\tau(x-\mu)/\sigma > 0 \qquad (1)$$

By using the (1) we define the response curve of the GEV regression

$$\pi(x) = \exp\{-[1+\tau(\alpha+\beta x)]^{-1/\tau}\}. \qquad (2)$$

For $\tau \to 0$ the model (2) becomes the Gumbel response curve and for $\tau > 0$ the Weibull response curve. The link function for this GLM is

$$\left\{[-ln\pi(x)]^{-\tau} - 1\right\}/\tau = \alpha + \beta x.$$

Let $Y_1, Y_2, ..., Y_n$ be a random sample, the log-likelihood and the score functions are

$$l(\beta,\tau,\mathbf{y}) = \sum_{i=1}^{n}\left\{-y_i[1+\tau\beta'\mathbf{x}_i]^{-1/\tau} + (1-y_i)ln[1-\exp\{-[1+\tau(\beta'\mathbf{x}_i)]^{-1/\tau}\}]\right\}$$

$$\frac{\partial l(\beta,\tau,\mathbf{y})}{\partial \beta_j} = -\sum_{i=1}^{n} x_{ij}\frac{\ln[\pi(\mathbf{x}_i)]}{1+\tau\beta'\mathbf{x}_i}\frac{y_i-\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} \quad j = 0,1,...,k,$$

$$\frac{\partial l(\beta,\tau,\mathbf{y})}{\partial \tau} = -\frac{1}{\tau^2}\sum_{i=1}^{n}\beta'\mathbf{x}_i ln[1+\tau\beta'\mathbf{x}_i]ln[\pi(\mathbf{x}_i)]\frac{y_i-\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}.$$

In order to estimate the parameters by the Newton algorithm, we define the initial values β^* and τ^*. We propose to set $\tau^* \simeq 0$, $\beta_j^* = 0\ \forall j$ and $\beta_0^* = \ln[-ln(\bar{y})]$, obtained by the Gumbel regression. Afterwards, by substituting β_0^* and $\beta_j^* = 0$ in the $\partial l(\beta,\tau,\mathbf{y})/\partial\tau = 0$ we obtaining the τ estimate for the first step of the iterative procedure. By using such estimate of τ and β_j^* in the $\partial l(\beta,\tau,\mathbf{y})/\partial\beta_j = 0$, we obtain the estimates of β_j with $j = 0,1,...,k$ for the first step in the GEV model.

The GEV regression is proposed to model a binary dependent variable that represents a rare event. Since defaults in credit risk analysis are rare events, we apply the GEV model to empirical data on Italian firms to model the default probability as a function of some covariates. The application shows that the logistic regression model underestimates the default probability. On the contrary, the GEV model overcomes this problem and accommodates skewness.

References

Agresti A. (2002). *Categorical Data Analysis*. New York: Wiley.

Kotz S. and Nadarajah S. (2000). *Extreme Value Distributions. Theory and Applications*. London: Imperial Colleg Press.

Ensemble procedures for more accurate Regression Trees with Moderating Effects

Gianfranco Giordano, Massimo Aria

1 Introduction

Within several research fields (sociology, economics, demography and health), it is likely to deal with hierarchical structure phenomenon, with multi-level data: individual, familiar, territorial and social. In such circumstances it is necessary to proceed with the analysis of the relation between individuals and the society, where naturally, can be observed at different hierarchical levels, and variables may be defined at each level. This leads to research into the interaction between variables characterizing individuals and variables characterizing groups. The measurement of this interaction has been defined "moderating effect".

This has been carried out by considering a non-parametric regression analysis (Giordano and Aria, 2010), that is based on a generalization of Classification and Regression Trees algorithm (Breiman et al., 1984) that takes into account the different role played by variables belonging to higher levels.

This paper points out how ensemble procedure in a regression tree methodology can be implemented that considers the relationships among variables belonging to different levels of a data matrix which is characterized by a hierarchical structure.

Starting from this approach, the problem of the robustness of the method are showed by implementing an ensemble procedure which focus its attention on the boosting philosophy (Freund and Schapire, 1997), trying to stress the potentialities and the advantages of such non parametric techniques.

Gianfranco Giordano,
Department of Mathematics and Statistics, University of Naples Federico II, Via Cintia, M.te Sant'Angelo, 80126 Napoli, e-mail: gianfranco.giordano@unina.it

Massimo Aria
Department of Mathematics and Statistics, University of Naples Federico II, Via Cintia, M.te Sant'Angelo, 80126 Napoli, e-mail: aria@unina.it

2 Main results

The Hierarchical dataset, as non-standardized data structures, are analyzed through Multilevel models. Nowadays, these statistical methodologies are considered more adequate to better extract the information of typical hierarchical structures. Like the classical approaches, multilevel analysis involves a wide range of theoretical assumptions and an equal range of problems for the interpretation of the results. In addition, when there are a lot of variables present at the different levels, there is a large amount of possible cross level interactions making it hard to estimate and interpolate the parameters (Snijders, 1999). A new approach called RTME (*Regression Trees with Moderating Effects*) has already been used, to deal with those limitations. This methodology is based on the generalization of CART partitioning criteria through the definition of a splitting algorithms that take into account the main moderating effect of the Z variable with respect to the prediction of X over Y. This is implemented on *Tree Harvest* software (Siciliano et al., 2004). This approach overcome the limitations of the CART methodology and furthermore, when in presence of situations where the functional and distributional assumptions of the multilevel model are not verified, or when its estimation algorithm does not converge, it is possible to use this technique to provide a viable and feasible alternative.

A well-known problem in literature concerning the tree-based method is its instability. To overcome this matter we propose, an ensemble procedure based on the boosting philosophy, that by using as weak-learner the RTME, allows to obtain a more accurate and robust regression. To use such a procedure, an ad-hoc error measure has been created, taking into account the moderating effect. This methodology has been tested on a great number of simulated and real datasets, underlining the characteristics of accuracy and robustness rather than the weak procedure.

References

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Freund Y., Schapire R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1).

Giordano G., Aria M., (2010). Regression Trees with Moderating Effects. Springer series in *New Perspectives in Statistical Modeling and Analysis*, forthcoming.

Siciliano R., Aria M., Conversano, C. (2004). Harvesting trees: methods, software and applications. In Proceedings in Computational Statistics: 16th Symposium of IASC Held in Prague, August 23-27. 2004 (COMPSTAT 2004), Eletronical Edition (CD) Physica-Verlag, Heidelberg.

Snijders T., Bosker R., (1999). *Multilevel Analysis. An Introduction to basic and advanced mutilevel modeling*, SAGE Publications, London.

On the use of bootstrap in factor analysis

John Ohrvik, Gabriella Schoier

1 Introduction

Different methods for variable reduction have been proposed (see Schoier , 2009), among these factor analysis is a multivariate correlation method that seek to explain the relation between a set of variables and smaller sets of underlying variables i.e. factors of potential importance by reducing a set of directly measured variables into a smaller set of underlying factors representing unique statistically uncorrelated domains termed factors.This may be useful in many different application fields.

2 Main results

In this work factor analysis is used to identify the initial set of uncorrelated components. The scree-test applying nonparametric bootstrap is used to determine the number of components to retain. It is based on the graph of descending eigenvalues λ_i of the correlation matrix of the input variables versus i and select as the number of components to retain the value of i corresponding to an 'elbow' in the curve, this is considered to be the point where large eigenvalues cease and small begin. To simplify the interpretation of the selected components the varimax (orthogonal) rotation was applied on the eigenvectors corresponding to the retained components. This has as its rationale the provision of uncorrelated factors with a few large loadings (= square root of eigenvalue times corresponding eigenvector, i.e. the matrix

John Ohrvik,
Karolinska Institutet, Department of Medicine, N3:06, SE-17176 Stockholm, Sweden,
e-mail: john.ohrvik@ki.se

Gabriella Schoier,
Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Universitá di Trieste,
Piazzale Europa 1, IT-34127 Trieste, Italia e-mail: gabriella.schoier@econ.units.it

of loadings times its transpose equals the correlation matrix) and as many near-zero loadings as possible. The proportion of bootstrap replicates with a different order of eigenvectors compared to that in the total sample was used as a measure of reliability of the factor analysis. These replicates were ignored when calculating bootstrap confidence intervals for the loadings, which were calculated using Efron's percentile method.

The results from the bootstrapped factor analysis were compared with those from 10-fold cross-validation2 of the factor analysis.

As application we performed a factor analysis of the continuous components of the Metabolic Syndrome (MetS) in a cohort of 196 men and 200 women comprising 65 % of a random sample of all 75-year-olds from the city of Vasteras in Sweden 1997 (see Ohrvik et al., 2009). Skewed variables (triglycerides, fasting glucose) were log-transformed and HDL-cholesterol, where high values are protective, was inverted prior to analysis.

The MetS consistently comprised two factors applying nonparametric bootstrap. Factor 1; a metabolic factor, consisted of fasting glucose (1), HDL-cholesterol (2), triglycerides (3) and waist (4). Factor 2; a blood pressure factor, consisted of diastolic (5) and systolic (6) blood pressure. These two factors explained 57.9% of the total variation; 1st factor eigenvalue 95% CI: 1.74-2.15 and 2nd factor 95% eigenvalue CI: 1.40-1.71. T he corresponding figures for women were 63.0%; 1st factor 2.00-2.43 and 2nd factor 1.43-1.72. For women the factor loadings showed consistent patterns over all the 10 000 bootstrap replicates while for men the 1st and 2nd factor were interchanged in 2.5% of the replicates. The results using bootstrap methods were consistent with those applying 10-fold cross-validation.

Applying bootstrap to factor analysis showed to be useful in selecting the number of factors to retain and to assess the consistency and accuracy of the factor analysis. Confidence intervals for the eigenvalues and factor loadings could be derived in a strait forward way applying Efron's percentile method after considering that the eigenvector space could change. Factor analysis of the basic variables of the Metabolic Syndrome among 75-year-olds from a general population identified a metabolic and a blood pressure factor.

References

Ohrvik J., Hedberg P., Jonason T., Lonnberg I., Nilsson G. (2009). Factor analysis of the individual components of the metabolic syndrome among elderly identifies two factors with different survival patterns - a population based study. *Metabolic Syndrome and Related Disorders*, 431-497.

Schoier G. (2009). On the problem of variable reduction in a customer satisfaction contexts. In: Classification and Data Analysis 2009, Book of Short Papers, Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, 621-624.

Contributed Session 16

Biomedical Applications

Evolutionary Neural Networks to design synthetic proteins

Davide De March, Irene Poli

1 Introduction

Natural proteins are the fundaments of everyday life activities but just a very small number of them are known. Since the number of all the possible proteins is potentially infinite (in fact the number of all possible combination of 20 amino acids in a protein with length l is 20^l), we are interested to know if and why the evolution selected only a very small number of proteins. The Darwinian process is generally accepted as an optimal selective process and this should demonstrate that the natural protein are the best proteins for our life activities.

However, some fundamental questions arise: are there other proteins (not natural) that might be able to have some good functionalities? Can a random protein become an existing one after the evolutionary process? (Luisi, 2003)

In this biological backgrounds, we deal with applications in which the number of variables p is much larger than the number of observations n (i.e when $n \ll p$). Such high dimensionality and these new scientific problems create great opportunities and significant challenges for the development of mathematical and statistical procedures. We are developing ensemble methods that combines the evolution strategy with the information extraction achieved by the statistical models (De March et al., 2009). The experimentation which this work is related to, deals with the detection of a natural protein that is hidden by the experimentalists in a sort of primordial broth.

Davide De March,
Department of Statistics at the University of Florence, e-mail: demarch@ds.unifi.it

Irene Poli,
Department of Statistics at the University of Venice e-mail: irenpoli@unive.it

2 Main results

The search in high dimensional space of proteins is realized assuming to decompose a catalytic protein of length 200 amino acids, in 4 strings of 50 elements (called domain). As a consequence, we determine 4 different positions in which the domains can be assigned. The biologists create 95 possible domains for each position and they hide the 4 target domains among them. The aim of the research is to detect the right domains in the right positions with the lowest number of experimental trials and to identify some "not-known proteins" that have a close functionality to the real one. The experimental space is extremely high ($8.15x10^7$), and the number of feasible real experiments is very small (about 500 trials in 5 experimental generations).

To simulate the natural darwinian process we adopt a Genetic Algorithm with a dynamic selection procedure, with a 5% of mutation rate and 1% of innovation and one point cross-over, and it performs very well in this particular setting (Goldberg, 1989).

We propose a novel approach to search in high dimensional space: the Evolutionary Neural Network. This approach is able to capture the most relevant information of the system in a very small set of data and to use it as a guide to design the following generation. Its outperforming capability is strictly connected with its main feature, that is its adaptability in topology. The procedure start with a very simple family of neural networks, and the network with the lowest prediction error is used to define the new generation of experimental points, achieving higher information about the system. Iterating this procedure for few generations the best Neural Network becomes more complex, describing relevant interactions among variables and embodying more and more information.

References

De March D., Slanzi D., Poli I. (2009) Evolutionary Algorithm for Complex Experimental Designs, in Ermakov S.M., Melas V.B., Pepelyshev A.N. (eds) *Proceedings of 6th St. Petersburg Workshop on Simulation 2009*: 547-551.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading,MA:Addison-Wesley.

Luisi, P. L. (2003) Contingency and determinism. *Philos. Transact A Math Phys Eng Sci*, 361:1141-1147.

Understanding co-expression of co-located genes using a PCA approach

Marion Ouedraogo, Frederic Lecerf, Sébastien Lê

1 Introduction

Studying the genome structure and its role in the gene function regulation could reveal new insights in the relationships between the regulation of the genes expression and their chromosomal locations (Chandra Janga et al., 2008).

The genome is distributed on several chromosomes where the locations of the genes could be interpreted as a spatial organization. Therefore, the main hypothesis is that some co-located genes could be involved in a common regulation due to either regulatory element or structural conformation. (Madan Babu et al., 2008).

At present, there is no method to identify the co-expression of several co-located genes at a genomic scale, so as to assess the role of the genome structure on the regulation of gene expression.

2 Main results

To address this problem we introduce the so called "autovariogram" (Figure 2), in reference to the autocorrelogram used in time series and the variogram used in spatial analysis, for understanding the seasonal and spatial dependencies respectively.

This autovariogram is obtained by means of a sequence of Principal Component Analysis (PCA) performed on sets of "co-located" genes; in other words on sets of consecutive genes. It displays graphically the variance of the first principal compo-

Ouedraogo Marion,
INRA, e-mail: marion.ouedraogo@rennes.inra.fr

Lecerf Frederic,
Agrocampus Ouest, e-mail: lecerf@agrocampus-ouest.fr

Sébastien Lê,
Agrocampus Ouest, e-mail: sebastien.le@agrocampus-ouest.fr

nent for consecutive sets of consecutives genes.This variance may be interpreted as co-expression.

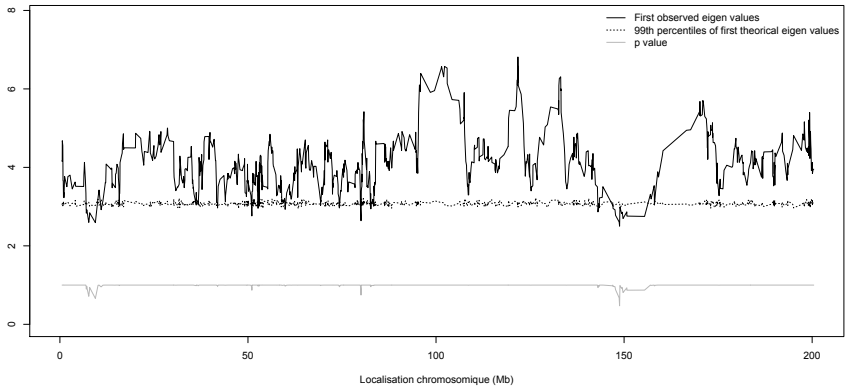


Fig. 1 Autovariogram of a chromosome. Firsts observed and simulated eigen values and p-values are plotted for each sets of genes according to their chromosomal location (median in Megabase).

This graphical representation is enhanced by p-values that corresponds each to the following test for a given set of genes.
H_0: The genes are not co-expressed.
H_1: The genes are co-expressed.
The aim of this talk is to illustrate the interest of the autovariogram and to presents some of its main features.

References

Madan Babu M., Chandra Janga S., de Santiago I., Pombo A. (2008). *Eukaryotic gene regulation in three dimensions and its impact on genome evolution*, Current Opinion in Genetics & Development, 18, 571–582.

Chandra Janga S., Collado-Vides J., Madan Babu M. (2008). *Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes*, Proceedings of the National Academy of Sciences, 41, 15761–15766.

Modeling delay to diagnosis for Amiotrophic Lateral Sclerosis: under reporting and incidence estimates

Irene Rocchetti

1 Introduction

The purpose of this work is to estimate the incidence rate for ALS (Amiotrophic Lateral Sclerosis) during the period 2007-2009; observed/available data are drawn from the National Registry of Rare Diseases (NRRD) which contains social, demographic and clinical information about a set rare diseases.

A subset of the subjects having a rare disease (in this context ALS) may have not been registered due to time/geographical coverage or underreporting and thus the corresponding number needs to be estimated for epidemiological purposes.

We propose to use an Horvitz-Thompson (HT) based approach to estimate the number of units suffering from ALS who have not been observed by the NRRD, to achieve an approximate incidence rate for ALS for years 2007-2009. The HT estimator helps correct the crude (registered) incidence rate by taking into account a standard, common, distribution for the delay to diagnosis.

We consider a KM estimator of the individual survivor function, representing the probability to have the disease diagnosed after a given time, which represents the probability of not having been registered by the Registry before that time. This probability is projected from older cohorts to younger ones to account for potential under reporting of cases with a more recent onset. The KM approach has been applied to the whole ALS sample after stratifying by some potential risk factors such as gender.

Irene Rocchetti,
Università La Sapienza, e-mail: irene.rocchetti@uniroma1.it

2 Main results

Registries can be seen as endogenous mechanisms which identify n units from a population of (unknown) size N (in this context the "real" ALS prevalent/incident population). The whole population size is given by the n sampled units and a part n_0 of units which has not been registered by the mechanism/s ($N = n + n_0$).

The maximum likelihood estimator of N is known to be the integer part of the Horvitz-Thompson estimator

$$\hat{N} = \left\lfloor \frac{n}{(1-p_0)} \right\rfloor$$

where p_0 is the probability of not registering the unit. We use a survival model to estimate p_0 and describe the delay to ALS diagnosis adopting some assumptions.

Let T denote the r.v "time to diagnosis" and t_i be the corresponding value for the i−th individual; furthermore, let x_i be the time of disease onset for the i−th individual. Considering all the registered cases, we are able to estimate the function $\Pr(Y \leq t)$ where $Y = T - x$ is the time from onset to diagnosis. Now, let t_{end} denote the end of the observation period (say the upper bound of the time period covered by the registry, in the present case, 21 december 2009). We assume that the i-th subject has not been registered if the corresponding diagnosis time is greater than t_{end}, thus

$$p_{0i} = \Pr(t_i > t_{end}) = \Pr[t_i - x_i > t_{end} - x_i] = \Pr[Y_i > (t_{end} - x_i)] = S(t_{end} - x_i)$$

The (extended) version of the Horvitz-Thompson (HT) estimator is as follows

$$N = \sum_{i=1}^{n} \frac{n_i}{1 - \hat{p}_{0i}}$$

where n_i identifies all units with the same onset time (in general $n_i = 1$). By using the survivor function estimated on the cohort 2001-2003 to weight the observed cases with onset in 2007-2009, the extended version of HT estimator provides $\hat{N} = 1295$ incident units having the ALS disease, while the estimate obtained by considering the survivor function estimated on the 2004-2006 cohort is $\hat{N} = 1123$.

References

Antolini L., Biganzoli, E.M., Boracchi, P. (2006). Crude Cumulative Incidence in the form of a Horvitz-Thompson like and Kaplan-Meier Estimator. *COBRA Preprint Series*.

Esbjerg S., Keiding, N., Koch-Henriksen, N. (1999). Reporting delay and corrected incidence of multiple sclerosis. *Statistics in medicine*, 18, 1691-1706.

Jisheng S. C., Yip, P. S. F., Chau, P.H. (2004). Estimation of reporting delay and suicide incidence in Hong Kong *Statistics in Medicine*, 23, 467-476.

Contributed Session 17

Developments in Archetypal Analysis

Archetypal Analysis for interval valued data

Stefania Corsaro, Marina Marino

1 Introduction

The use of interval data instead of single valued data allows to take into account valuable information, as variability and/or uncertainty which may be inherent to the data, that otherwise should be lost. In this talk, we presented an extension of the model for Archetypal Analysis of real numbers proposed by Cutler and Breiman (1994) to handle real interval numbers.

In order to develop our model, we will refer to the matrix formulation of the Archetypal Analysis problem. Let us organize the data into a matrix $X = (x_{ij}) \in \Re^{m \times n}$, in which the rows refer to the individuals and the columns to the variables. Then, given an integer p, the core problem of Archetypal Analysis is to find matrices $A = (\alpha_{ik}) \in \Re^{m \times p}$, $B = (\beta_{ki}) \in \Re^{p \times m}$ which solve the non-convex minimization problem:

$$\min_{A,B} f(A,B) = \min_{A,B} \|X - A \cdot B \cdot X\|_F, \tag{1}$$

where $\|\cdot\|_F$ denotes, as usual, the Frobenius norm, under the constraints

$$\begin{array}{ll} \alpha_{ik} \geq 0; & \sum_{k=1}^{p} \alpha_{ik} = 1, \quad i = 1, \ldots, m; \\ \beta_{ki} \geq 0; & \sum_{i=1}^{m} \beta_{ki} = 1, \quad k = 1, \ldots, p. \end{array} \tag{2}$$

The archetypes are then the rows of the matrix $Z \in \Re^{p \times n}$ given by the product $Z = B \cdot X$.

Stefania Corsaro,
Dept. Statistics and Mathematics for Economic Research, University of Naples Parthenope,
e-mail: stefania.corsaro@uniparthenope.it

Marina Marino,
Dept. Mathematics and Statistics, University of Naples Federico II,
e-mail: marina.marino@unina.it

2 Main results

Let $\mathbf{A}$, $\mathbf{B}$ be two interval matrices. We define *distance matrix* between $\mathbf{A}$ and $\mathbf{B}$ the non-negative matrix $(q(\mathbf{A},\mathbf{B})_{i,j}) := (q(\mathbf{A}_{ij},\mathbf{B}_{ij}))$, that is, the pointwise distance, to be meant in the sense of Hausdorff distance, between the elements of $\mathbf{A}$ and $\mathbf{B}$. It can be shown that if $||\cdot||$ denotes a real matrix norm, then $||q(\mathbf{A},\mathbf{B})||$ defines a metric on the set of interval matrices (Alefed , 1990).

The Interval Archetypal Analysis (IAA) problem can be stated in a similar way as problem for AA for real data. Let $\mathbf{X} \in \mathfrak{I}\Re^{m\times n}$ be an interval matrix, in which the rows represent the individuals and the columns the variables. Given an integer p, the Interval Archetypal Analysis (IAA) aims at finding matrices $A = (\alpha_{ik}) \in \Re^{m\times p}$, $B = (\beta_{ki}) \in \Re^{p\times m}$ which solve the non-convex minimization problem:

$$\min_{A,B} \|q(\mathbf{X}, A\cdot B\cdot \mathbf{X})\|_F = \|\,|X_c - (A\cdot B\cdot \mathbf{X})_c| + |\Delta X - \Delta(A\cdot B\cdot \mathbf{X})|\,\|_F, \qquad (3)$$

under the constraints (2). In (3), X_c and ΔX denote the *center* and the *radius matrices* respectively.

The choice of metric in (3) is motivated by the fact that this metric allows us to keep under control both the distance between the centers, for the sake of localization, and the width of the involved intervals, for the sake of accuracy. Moreover, the Frobenius norm is the natural choice for it is the one employed in the original definition of Archetypal Analysis given by Cutler and Breiman, and it can be efficiently computed.

The archetypes are then given by the row vectors of the matrix $\mathbf{Z} \in \mathfrak{I}\Re^{p\times n}$ defined by the product $\mathbf{Z} = B\cdot\mathbf{X}$.

The IAA problem is again a non-convex, single-valued optimization problem.

To validate our model and, mainly, to analyze the coherence between the results obtained from Archetypal Analysis of single-valued data and the ones obtained working with interval data, we developed our procedures in MatLab environment, using a routine of MatLab Optimization Toolbox (2008) to solve the core minimization problem both in the single-valued case and in the interval data case. Results, that are presented and extensively discussed in Corsaro and Marino (to appear), seem promising, since the interval-based model produces results which are as accurate as those obtained in the single-valued data case.

References

Alefeld G., Herzberger J. (1990). *Introduction to Interval Computations*. Academic Press, NY.

Corsaro S., Marino M. (to appear). *Archetypal Analysis of interval data*. Reliable Computing.

Cutler A., Breiman L. (1994). Archetypal analysis. *Technometrics*, 36, 338-347.

MatLab Optimization Toolbox 4 (2008). *User's Guide*. The MathWorks, Inc.

Archetypal analysis for prototype identification

Maria Rosaria D'Esposito, Francesco Palumbo, Giancarlo Ragozini

1 Introduction

In many real application contexts, clustering is an important learning technique widely used to identify different typologies in the data. To represent each cluster, it may turn out useful to have a single *datum*, which is called *prototype*. However, assuming a single unit as representative of a cluster may be to limited and produce an heavy loss of information. Instead of a single unit, some authors have proposed to consider as prototype a set of units or an interval valued unit. Many contributes came from Symbolic Data Analysis literature, with respect to the clustering of interval valued data (Diday and Noirhomme-Fraiture, 2008).

The most known approach to obtain prototypes, in the statistical clustering framework, certainly is constant radius method (such as the k-means algorithm and the related *moving center methods*). Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, $\mathbf{X} \in \Re^{n \times p}$, moving center methods aim at minimising the function $J_K = \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} d(\mathbf{v}_k, \mathbf{x}_{i_k})^2$ with respect the K prototypes $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k\} \in \Re^{k \times p}$: where K indicates the number of cluster that is *a priori* fixed, $\{n_1, n_2, \ldots, n_k\}$ indicate the number of units in each cluster, and $d(\cdot, \cdot)$ is a suitable distance function. Moving center methods assure good results when dealing with elliptical clusters. In other cases the method can become unstable and can not allow a correct clusters identification.

The present paper proposes a *new* method for the prototype identification based on the archetypes (Cutler and Breiman, 1994), "pure individual types", few points lying on the boundary of the data scatter and characterizing the archetypal pattern in the data.

Maria Rosaria D'Esposito,
Dept. of Economics and Statistics, e-mail: mdesposito@unisa.it

Francesco Palumbo,
Dept. of Relational Science,

Giancarlo Ragozini,
Dept. of Sociology, e-mail: giragoz@unina.it

2 Main results

The procedure to identify the prototypes is based on geometrical and statistical properties of archetypes. Given the data matrix $\mathbf{X}$, both single valued or interval valued, the archetype matrix $\mathbf{A}(m) = (\mathbf{a}_1, \ldots, \mathbf{a}_m)'$, $\mathbf{A}(m) \in \Re^{m \times p}$, and convex combination coefficients $\beta(m) = (\beta_1, \ldots, \beta_m)$ and $\alpha(m) = (\alpha_1, \ldots, \alpha_n)'$, the archetypes are those m points minimizing $\min_{\alpha(m),\beta(m)} \|\mathbf{X} - \alpha(m)\beta'(m)\mathbf{X}\|_F$ where $\|\mathbf{Y}\|_F = \sqrt{Tr(\mathbf{Y}\mathbf{Y}')}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$.

The archetypes are vertices of a simplex in the data space $\Re^p$, and for each data point $\mathbf{x}_i'$ new coordinates $(\lambda_{i1}, \ldots, \lambda_{im})$ in a space spanned by the archetypes can be obtained by solving the equation $(\lambda_{i1} + \cdots + \lambda_{im})\,\mathbf{x}_i' = \lambda_{i1}\,\mathbf{a}_1' + \cdots + \lambda_{im}\,\mathbf{a}_m'$. $(\lambda_{i1}, \ldots, \lambda_{im})$ are the so called barycentric coordinates, and coincide with the $\alpha_{ij}(m)$ coefficients (Porzio et al., 2008). The associate space allows to explore and analyze data from an outward-inward perspective, and in it we propose to use the archetypes as starting points for a k-mean type clustering procedure with a number of clusters equal to the number of archetypes. As the α's coefficients can be viewed as compositional data, in the clustering procedure we use a proper compositional distance (Aitchison et al., 2000) Given the clusters so constructed, we can define prototypes in terms of single units, set of units or interval valued units.

Archetypes analysis presents many advantages for the prototype identification: *i*) the number of archetypes can be determined through an optimization criterion confronted with a certain degree of arbitrariness of moving center method; *ii*) clustering is performed in the space spanned by the archetypes, reducing computational costs and allowing to treat different type of data; *iii*) archetypes are extreme points and clustering based on them yields prototypes will be well-separated and with clear profiles that can be interpreted through the graphical and exploratory tools proposed for the archetypal analysis (D'Esposito et al., 2010).

References

Aitchison J., Barceló-Vidal C., Martýn-Fernández J.A., Pawlowsky-Glahn V. (2000). Logratio Analysis and Compositional Distance. *Mathematical Geology*, 32, 271–275.

Cutler A., Breiman L. (1994). Archetypal Analysis. *Technometrics*, 36, 338–347.

D'Esposito M.R., Ragozini G., Vistocco D. (2010). Exploring Data through Archetypes. In: Classification as a Tool for Research, Locarek-Junge H. and Weihs C. EDS. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin: Springer *in press*.

Diday E., Noirhomme-Fraiture M. EDS (2008). *Symbolic Data Analysis and the SODAS Software*, Chichester: Wiley.

Porzio G.C., Ragozini G. Vistocco D. (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry*, 24, 419–437.

New features for archetypal analysis in R

Manuel Eugster, Friedrich Leisch

1 Introduction

Archetypal analysis (Cutler and Breiman, 1994) has the aim to represent observations in a multivariate data set as convex combinations of a few, not necessarily observed, extremal points (archetypes). The archetypes themselves are restricted to being convex combinations of the individuals in the data set and lie on the boundary of the data set.

Archetypal analysis approximates the convex hull of the data set – this suggests itself that outliers have a great influence on the solution. In fact, the farther a data point is from the center of the data set, the more influence it has on the solution. Although archetypal analysis is about the data set boundary, practice has shown that in many cases one primarily is interested in the archetypes of the large majority than of the totality.

2 Main results

R package archetypes (Eugster and Leisch, 2009) is a new freely available and very flexible implementation of the algorithm, that can easily be modified. We have adapted the original archetypes estimator to be a robust M-estimator and present an iteratively reweighted least squares fitting algorithm (Eugster and Leisch, 2010). Our robust archetypal analysis algorithm is based on weighting the residuals and ob-

Manuel Eugster,
Department of Statistics, Ludwig-Maximilians-Universität München,
e-mail: Manuel.Eugster@stat.uni-muenchen.de

Friedrich Leisch,
Department of Statistics, Ludwig-Maximilians-Universität München,
e-mail: Friedrich.Leisch@stat.uni-muenchen.de

servations, respectively. As a byproduct we hence obtain weighted archetypal analysis which enables us to represent additional information available from the data set, like the importance of observations or the correlation between observations.

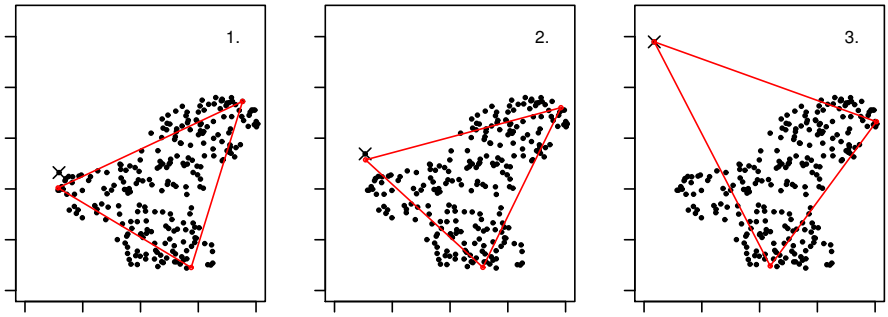


Fig. 1 Behavior of the archetypes (triangle) when one data point (unfilled dot) constantly moves away (cross).

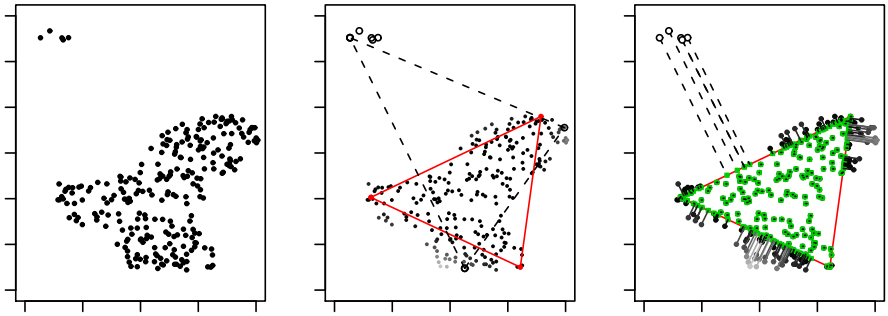


Fig. 2 Robust archetypal analysis; the gray scale of the data points indicate their final weights. Note that the outliers have weight 0 (unfilled).

References

Cutler A., Breiman L. (1994). Archetypal analysis. *Technometrics*, 36(4), 338–347.

Eugster M.J.A., Leisch F. (2009). From Spider-man to Hero - archetypal analysis in R. *Journal of Statistical Software*, 30(8), 1–23, URL http://www.jstatsoft.org/v30/i08

Eugster M.J.A., Leisch F. (2010) Weighted and robust archetypal analysis. Technical Report 82, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, URL http://epub.ub.uni-muenchen.de/11498/

Contributed Session 18

Model Selection

Automatic detection of outliers in linear regression models: the Forward Search approach

Bruno Bertaccini, Franco Polverini

This article was stimulated by the work of Atkinson and Riani (2000) on the estimation of regression models using a robust methodology called by the authors forward search, which seems to work well in the estimation of a variety of models, particularly when part of the data are generated by models different from the one we intend to estimate. The methodology detects the presumed outliers and allows the estimation of the model without them. The weak (perhaps strong) point of the procedure is, in our opinion, the fact that the choice of the final set of data to be used for the estimation is left to the researcher who relies on the behavior, often the visual behavior, of some statistics, as observations are added to an initial small subset of the data at hand. The choice is hence subjective and by its very nature impedes the use of simulations to judge how good the method is. The paper proposes some automatic ways of making this choice which allows the running of simulations in order to assess the properties of the estimators and make comparisons with alternative ways of estimating the models involved. We ran some simulations with the proposed methodology, computed efficiency of the estimators and compared it with OLS and LMS estimators.

References

Atkinson A.C., Riani M. (2000). *Robust Diagnostic Regression Analysis*, Springer Verlag, New York.

Bruno Bertaccini,
Department of Statistics “G. Parenti”, University of Florence, e-mail: bertaccini@ds.unifi.it

Franco Polverini,
Department of Statistics “G. Parenti”, University of Florence, e-mail: polverini@ds.unifi.it

Estimating the Gap of the Forward Search via censored sampling

Danya Facchinetti, Silvia Angela Osmetti

1 Introduction

The recent literature shows the importance of the robust methods used in presence of outliers in the inferential procedures. Instead the robust statistics provide powerful tools to detect outliers and to compute stable and efficient estimates. Atkinson, Riani and Cerioli (Atkinson et al., 2004) have developed the Forward Search (FS), a robust iterative technique for detecting multiple outliers and determining their effects on models fitted to the data.

The idea of this method is to identify the outliers starting from a small, robustly chosen, subset of the data and to move to a larger subsets by including, at each step, a new observation. The progress of the search depends on ordering of the all data through the Mahalanobis distance (MD) (De Maesschalck et al., 2000) and on selecting the observation that has small distance and that is so unlikely to be outlier. At the end of the iterative procedure we identify two groups: the potential outliers and the clean subsets of data.

The progression of the search is monitored by specific indicators, for example the Gap statistic that measures the difference among two consecutive Mahalanobis distances. The Gap is used in the forward search to identify the outliers in a data set. In this paper, under the assumption of absence of outliers and of multinormal distribution law, we study the well known distribution law of the "Gap" statistic. In particular, we propose to estimate the parameters of the Mahalanobis distances, considering the II Type censored sampling technique (Cohen, 1991). In such way we considers all the observations of the data-set for the parameters estimate. On the contrary the Forward Search uses only the observation inside the subsets. The results are verified through several simulation studies.

Danya Facchinetti, Silvia Angela Osmetti,
Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore,
e-mail: danya.facchinetti@unicatt.it, silvia.osmetti@unicatt.it

2 Main results

Let **x** be a data set of the n observations of a p-dimensional numerical vector. Let μ be the $p \times 1$ vector of the means and Σ be the $p \times p$ covariance matrix of the data. The MD d^2 is

$$d^2 = (\mathbf{x} - \mu)'(\Sigma)^{-1}(\mathbf{x} - \mu).$$

In the FS the two unknown parameters (μ, Σ) are estimate at each step of the procedure. The FS estimates the parameters by the observations into the outlier free subset of $m < n$ size. Alternatively we consider a censored sample of the data.

Let m_0 be the size of the starting subset, that is free of outliers, the mean and covariance matrix estimates are $\widehat{\mu}_0$ and $\widehat{\Sigma}_0$. From these parameter estimates we calculate and ordinate a set of n d^2:

$$d^2_{(1)m_0}, d^2_{(n),m_0} ..., d^2_{(m_0),m_0}, d^2_{(m_0+1),m_0}, ..., d^2_{(n),m_0}.$$

We find a II type censored sample of the distances, stopping the experiment after the observation of $m_1 = m_0 + 1$ distances. The $d^2_{m_0+1}$ observed value is assigned to the all $n - m_1$ not observed distances:

$$d^2_{(1)m_0}, d^2_{(n),m_0} ..., d^2_{(m_0),m_0}, d^2_{(m_0+1),m_0}|, d^2_{(m_0+1),m_0}, ..., d^2_{(m_0+1),m_0}.$$

The n observations, corresponding to the distances of the censored sample, describe the new set free of outliers that is used to obtain the new parameter estimates $\widehat{\mu}_1$ and $\widehat{\sigma}_1$ at the first step. On the contrary the classic FS used only the $m_0 + 1$ observation. The iterative procedures improve, as m goes from m_0 to n, until the outliers are identified. In the FS there are three aspects: starting, progressing and monitoring the search. Our proposal modify the progressing and so the monitoring of the search. The censored sample allow us to estimate the MD parameters using all the n informations known at the generic step m of the procedure. Since the FS depends on ordering of the n distances, therefore seem more coherent to estimate the parameters using this type of sampling.

In the FS the outlier detection is based on the monitoring of quantities such as d^2, the Gap and other diagnostic quantities. We are interested to evaluate the ability of outlier detection using these estimate from censored sample and the consequences on diagnostic quantities, in particular on the Gap and its distribution.

References

Atkinson A. C., Riani M. , Cerioli A.(2004). *Exploring Multivariate Data with the Forward Search* . New York: Springer Verlag.

Cohen C. A. (1991). *Truncated and censored samples*. New York: Marcel Dekker.

De Maesschalck R., Jouan-Rimbaud D., Massart D.L. (2000).The Mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50, 1-18.

Assessing model accuracy using a two-dimensional loss function

Roberto Savona, Marika Vezzoli

1 Introduction

In this paper we introduce a global accuracy measure for statistical models when jointly considering *in-* and *out-of-sample* accuracy. We do this by introducing a two-dimensional Loss Function (LF) computing a weighted average of in- and out-of-sample LF with weights reflecting the decision-maker's objective (data generating process vs. forecasting activity). Non-parametric statistics and stochastic dominance criteria are used to get an objective evaluation of the models using subjective preferences. In other terms, starting from subjective evaluations about in- and out-of-sample model reliability, we come to select the best model by averaging fitting and predicting ability together with low and high risk aversion.

2 Main results

The models accuracy is assessed in-sample and out-of-sample through the LF derived from the ROC curve and its summary measure, the Youden Index (YI), in order to maximize the overall classification ability, minimizing both type-I (α) and type-II (β) errors. Let λ be the cut-off point of the YI. Formally, the YI is

$$\mathrm{YI} = \max_{\lambda} \left\{ [1 - \alpha(\lambda)] + [1 - \beta(\lambda)] - 1 \right\}. \tag{1}$$

Denoting with λ_{YI} the optimal threshold value obtained maximizing YI on λ, the LF is computed as the weighted sum of α and β with cost ζ and $(1-\zeta)$, respectively

$$\mathrm{LF} = [\zeta \cdot \alpha(\lambda_{\mathrm{YI}}) + (1-\zeta) \cdot \beta(\lambda_{\mathrm{YI}})]. \tag{2}$$

Roberto Savona and Marika Vezzoli
University of Brescia, Department of Business Studies and Department of Quantitative Methods, C.da Santa Chiara 50, 25122 Brescia, e-mail: savona@eco.unibs.it,vezzoli@eco.unibs.it.

Hence, letting by ω and $(1-\omega)$ the weights for in- and out-of-sample errors and starting from equation (2), the two dimensional LF (2^{D}LF) that we propose becomes

$$2^{\mathrm{D}}\mathrm{LF} = \zeta \cdot \left[\omega \cdot (\alpha(\lambda_{\mathrm{YI}}))^{\mathrm{in}} + (1-\omega) \cdot (\alpha(\lambda_{\mathrm{YI}}))^{\mathrm{out}}\right] + \tag{3}$$

$$+ (1-\zeta) \cdot \left[\omega \cdot (\beta(\lambda_{\mathrm{YI}}))^{\mathrm{in}} + (1-\omega) \cdot (\beta(\lambda_{\mathrm{YI}}))^{\mathrm{out}}\right]$$

where $(\alpha(\lambda_{\mathrm{YI}}))^{\mathrm{in}}$ and $(\alpha(\lambda_{\mathrm{YI}}))^{\mathrm{out}}$ denote the type-I errors in- and out-of-sample, respectively, computed in correspondence of λ_{YI}. Analogously, $(\beta(\lambda_{\mathrm{YI}}))^{\mathrm{in}}$ and $(\beta(\lambda_{\mathrm{YI}}))^{\mathrm{out}}$ denote the type-II errors in- and out-of-sample relative to λ_{YI}.

Using annual observations for 66 emerging economies over the period 1975-2002 with 112 crisis episodes, of which 40 used in the out-of-sample analysis over the sub-period 1990-2002, we compute the 2^{D}LF for: (i) Logistic Regression (Logit); (ii) Logistic Stepwise Regression (sLogit); (iii) Noise-to-Signal Ratio (KLR) (Kaminsky et al., 1998); (iv) Regression Trees (RT); Final Model (FM) (Vezzoli and Stone, 2007). As shown in Figure 1, from modest to high risk aversion ($0.4 \leq \zeta \leq 0.8$) the FM appears to be the best model, as also confirmed by the median cost for the FM which is lesser than what shown by alternative models, also exhibiting low dispersion relative to competing models.

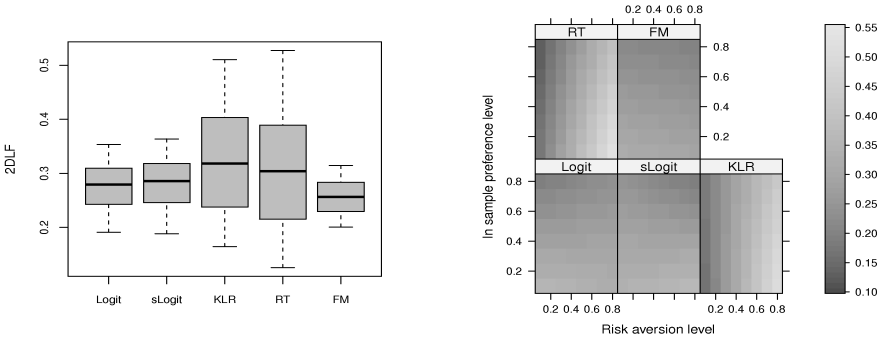


Fig. 1 Box plot and Contour plot for 2^{D}LF

References

Kaminsky G., Lizondo S., Reinhart C.M. (1998). Leading Indicators of Currency Crises. *Staff Papers - International Monetary Fund*, 45 (1), 1-48.

Vezzoli M., Stone C.J. (2007). CRAGGING. In *Book of Short Papers CLADAG 2007*, EUM, 363-366.

Contributed Session 19

Proximity Data and Hierarchies

Multi-objective genetic algorithm based clustering for dissimilarity data

Laura Bocci

1 Introduction

The clustering problem is defined as the problem of partitioning a collection of objects into a set of homogeneous clusters without any a priori knowledge. Generally, objects are described by J attributes and the number of classes is unknown but usually assumed to be much smaller than the number of objects. However, when only the observed pairwise dissimilarities of a set of objects are available, two common alternative approaches to partitioning objects from dissimilarities are used. In the first, a hierarchical clustering algorithm is applied and subsequently a partition is chosen by visual inspection of the resulting dendrogram. Alternatively, a tandem analysis is used by applying first a Multidimensional Scaling (MDS) method on dissimilarities and subsequently a partitioning algorithm to the low-dimensional point configuration resulting from MDS. Vichi (2008) proposed three clustering models for dissimilarity data following the model-based approach, which consists of fitting the "closest" classification matrix to the observed dissimilarity matrix $\mathbf{D}$.

Moreover, a fundamental difficulty of clustering is the determination of the number of clusters in the data set. Traditional clustering algorithms, in general, do not produce alternative solutions, and most of them do not lead to the optimal number of clusters in the dataset that they work on. Most existing algorithms require this number to be provided as a parameter which is a major problem in a setting where the structure of the data is completely unknown. Multi-objective genetic clustering technique (Handl and Knowles, 2007; Sriparna and Bandyopadhyay, 2010) has been proposed in the literature to simultaneously optimizes more than one objective function for automatically partitioning a set of objects described by J attributes.

In this paper, a multi-objective clustering technique is proposed to find the appropriate partition of a set of objects directly from dissimilarity data as well as automat-

Laura Bocci
Department of Sociology and Communication, Sapienza University of Rome, Via Salaria 113 - 00198 Roma, Italy, e-mail: laura.bocci@uniroma1.it

ically determine the proper number of clusters. We propose a clustering technique based on the explicit optimization of a partitioning with respect to multiple, complementary clustering objectives which take into account both heterogeneity for each class and isolation between classes, as described by Vichi (2008). The algorithm will automatically attain a set of Pareto-optimal solutions in terms both of the number of clusters and the appropriate classification of objects into groups.

2 Main results

We propose a multiobjective genetic clustering algorithm which uses two different validity criteria, the Variance Ratio Criterion (VRC) and the Davies-Bouldin (DB) index, as fitness functions. The chromosome for encoding a candidate solution is a string of fixed length n of integer values. If a chromosome has K clusters ($2 \leq K \leq K_{max}$), then each gene, which relates to a specific item, may take a value between 1 and K indicating the cluster to which the item itself belongs. Here, each chromosome corresponds to a partition C in K clusters of the set of n objects. For such a partition, both a square dissimilarity matrix $\mathbf{D}_B$ of order K and a diagonal matrix $\mathbf{D}_W$ of order K are estimated from $\mathbf{D}$ (Vichi, 2008). $\mathbf{D}_B$ consists of the $K(K-1)$ measures of isolation for pairs of clusters; while $\mathbf{D}_W$ has the measure of heterogeneity for each cluster as diagonal entries (Vichi, 2008). After, the two fitness function values of the chromosome are calculated on the basis of the elements of $\mathbf{D}_B$ and $\mathbf{D}_W$. In order to attain a good partitioning, the former validity criterion (VRC) should be maximized, while the latter (DB index) should be minimized. Therefore, these two objectives are contradictory in nature and can balance each other critically guiding the algorithm to global optima solutions. This allows us to achieve a set of partitionings showing different trade-offs between the two objectives and consisting of different numbers of clusters. The population is evolved by applying the genetic operators selection, one point crossover and mutation. Moreover, the elitist strategy is implemented so that at each generation the non-dominated solutions among the parent and child populations are propagated to the next generation. The proposed algorithm has been tested in a simulation study and with a data set analyzed in the literature.

References

Handl J., Knowles J. (2007). An Evolutionary Approach to Multiobjective Clustering. *IEEE Trans. on Evolutionary Computation*, 11, 56-76.

Sriparna S., Bandyopadhyay S. (2010). A symmetry based multiobjective clustering technique for automatic evolution of clusters. *Pattern Recognition*, 43, 738-751.

Vichi M. (2008). Fitting semiparametric clustering models to dissimilarity data. *Advances in Data Analysis and Classification*, 2 (2), 121-161.

A new approach for analyzing a set of hierarchies

Marine Cadoret, Sébastien Lê, Jérôme Pagès

1 Introduction

The aim of this presentation is to propose a new approach for analyzing hierarchies issued from unsupervised classifications performed on the same (statistical) individuals. This issue has already been partially addressed by several authors for the comparison of different classifications methods (Leclerc (1985); Leclerc and Cucumel (1987) or the special issue of *Journal of Classification* (Vol. 3, 1986) dedicated to the comparison and consensus of classifications).

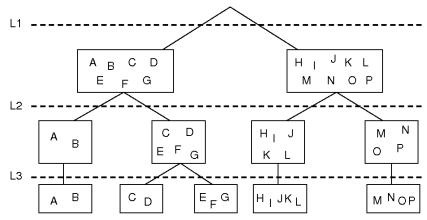
The starting point of our research framework is the hierarchical sorting task commonly used in psychology and sensory analysis. This method consists in asking subjects to provide each their own hierarchical tree from the same given set of objects. This hierarchical tree is constructed mostly in a binary and descending way: the subjects are asked to divide the objects into two homogeneous groups and then to divide again each of the two groups until they consider the final groups homogeneous. The main feature of this method is that each subject uses his/her own criteria for making these successive divisions. In this kind of experiment we're interested into getting a consensus representation of the objects from all the subjects as well as a representation of the subjects, function of the way they classified the objects.

In this talk, we propose a method which provides on the one hand a Euclidean representation of the objects and on the other hand a Euclidean representation of the hierarchies (i.e. a subject can be assimilated to his/her hierarchy) linked to the previous one in the manner of Multiple Factor Analysis (MFA, Escofier and Pagès (1998)). This hierarchy representation allows to visualize the different steps taken by each subject and to understand in a certain way his/her cognitive process.

Marine Cadoret, Sébastien Lê, Jérôme Pagès,
Agrocampus, Applied mathematics department, Rennes, France,
e-mail: marine.cadoret@agrocampus-ouest.fr

2 Main results

The data associated with a hierarchy j can be gathered in a data table with I rows and Q_j columns (with Q_j the number of levels associated with the hierarchy j). In the example of figure 1, the hierarchy is composed of three levels. This hierarchy is associated with the data table of figure 1 (right). This data table comprises the $I = 16$ objects in rows and the three levels in columns. Each level of the hierarchy corresponds to a qualitative variable with as many modalities as there are groups for this level.



	L1	L2	L3
A	1	1	1
B	1	1	1
C	1	2	2
D	1	2	2
E	1	2	3
F	1	2	3
G	1	2	3
H	2	3	4
I	2	3	4
J	2	3	4
K	2	3	4
L	2	3	4
M	2	4	5
N	2	4	5
O	2	4	5
P	2	4	5

Fig. 1 Example of a hierarchy and the associated data table.

The data coming from a set of hierarchies can be gathered in a table that juxtaposes the tables associated with each hierarchy. This data table is composed of I rows and $Q = \sum_j Q_j$ columns: each row corresponds to an object and each column to a level associated with a given hierarchy; the columns are grouped by hierarchy.

This kind of table only composed of qualitative variables structured in groups (one group = one hierarchy) can be analyzed by Multiple Factor Analysis (MFA): Escofier and Pagès (1998). MFA is applied to a data table in which the same set of individuals (here the objects) are described by several sets of variables (here the hierarchies) structured in groups. The MFA balances the influence of each group of variables (i.e. each hierarchy) in the analysis making the maximum axial inertia of the clouds associated with the separated analysis of each hierarchy equal to 1.

References

Escofier B., Pagès J. (1998). *Analyses factorielles simples et multiples*, Paris: Dunod.

Leclerc B. (1985). La comparaison des hiérarchies: indices et métriques. *Mathématiques et sciences humaines*, 92, 5-40.

Leclerc B., Cucumel G. (1987). Consensus en classification : une revue bibliographique. *Mathématiques et sciences humaines*, 100, 109-128.

Extensions of modularity clustering

Andreas Geyer-Schulz, Michael Ovelgönne

1 Introduction

The purpose of this paper is to extend the randomized greedy modularity clustering algorithm of Ovelgönne and Geyer-Schulz (2010) for weighted, asymmetric networks and to investigate variants of the modularity measure. This is motivated on the one hand by the recent result of the limited resolution of the modularity measure Fortunato and Barthélemy (2007) and on the other hand by the experiments of Brandes et al. (2008) with various cluster indices.

The modularity measure has been introduced by Newman as a measure of cluster quality for undirected unweighted graphs. The modularity measure compares the empirical distribution of observed in-cluster connections with the expected distribution of a randomly generated graph with the same degree distribution (see e.g. Newman (2004)). The expected distribution of a randomly generated graph with the same degree distribution acts as a global null model for assessing the quality of a graph partition.

2 Main results

Table 1 shows that the randomized greedy modularity clustering algorithm Ovelgönne and Geyer-Schulz (2010) which pursues a hierarchical agglomerative clustering strategy compares favorably on a set of 8 datasets with the previous best algorithm of Zhu et al. (2008) on undirected and unweighted graphs.

Andreas Geyer-Schulz,
KIT, Kaiserstraße, 12, D-76128 Karlsruhe, e-mail: Andreas.Geyer-schulz@kit.edu

Michael Ovelgönne,
KIT, Kaiserstraße, 12, D-76128 Karlsruhe, e-mail: Michael.Ovelgoenne@kit.edu

Table 1 Run-time und quality comparison of 100 test-runs.

Network			RG(1)-VM		MOME	
Name	Vertices	Edges	Q	Time	Q	Time
Karate	34	78	0.412	0.004	0.420	0.002
Jazz	198	2742	0.444	0.015	0.444	0.011
Email	1133	5451	0.572	0.038	0.573	0.082
PGP	10680	24340	0.880	0.242	0.880	0.748
Cond.Mat.	27519	116181	0.749	0.972	0.756	2.588
WWW	325729	1090108	0.936	10.178	0.937	28.68
Amazon	410236	3356824	0.885	24.5	0.861	46.2
LiveJournal	4843953	42845684	0.760	481.9	oom	oom

As implemented, the algorithm also works with undirected weighted networks. However, for this case the modularity measure must be slightly adjusted in its interpretation taking into account the weighting of the connections.

For weighted asymmetric networks, the abstract data type supporting fast access has been properly adapted and the computation of the delta of the modularity function has also been adapted. The adaptations slow down the algorithm, because the exploitation of the symmetry of the undirected case ceases to be possible. The modularity measure must now be interpreted with regard to the weighting. Computations of asymmetric weighted data sets are in preparation. In addition, other global and local null models will be investigated.

References

Brandes U., Gaertler M., Wagner D. (2008). Engineering Graph Clustering: Models and Experimental Evaluation. *Journal of Experimental Algorithmics*, Vol. 12, 1–26.

Fortunato S., Barthelemy M. (2007). Resolution Limit in Community Detection. *Proceedings of the National Acadamy of Sciences of the USA (PNAS)*, Vol. 104(1), 36-41.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*. Vol. 69, 066133.

Ovelgönne M., Geyer-Schulz A. (2010). Randomized Greedy Modularity Optimization for Group Detection in Huge Social Networks. In: The 4th SNA-KDD Workshop, ACM, (submitted).

Zhu Z., Wang C., Ma L., Pan Y., Ding Z. (2008). Scalable Community Discovery of Large Networks. In: *WAIM'08: Proceedings of the 9th International Conference on Web-Age Information Management*, 381-388.

Contributed Session 20

Professional Profiles

University graduate's job hunting: what helps to get the good one?

Marisa Civardi, Franca Crippa, Vincenzo Bagnardi

1 Introduction

The rate of graduates' entry in the labour market is the most direct indicator of the external efficacy of university education. More attention, though, has been recently payed to the adequacy of their job with respect to the intensity of use of the human capital accumulated at university and to the role in determining the educational effectiveness (Fabbris et al., 2010). So far, indicators of job quality have been typically expressed in terms of wage and other aspects, mainly defined by outside observers only rather than as experienced by workers themselves (Clark, 2004). Nine hundred and sixty one graduates in 2007, employed at the moment of the survey, were interviewed twelve months after graduation on the characteristics of the present occupation by STELLA (an acronym of Statistiche sul Tema Laureati e Lavoro in Archivio on line). They were asked both objective and subjective aspects of their present job and of their future labour market behaviour. Taking into account aspects often neglected allowed to reshape the overriding role of wage in a more comprehensive setting and to set graduates' transition to the job market in a wider spectrum of 'good' outcomes.

2 Main results

Several aspects of graduates' job were asked, both on the contract and on the adequacy of the job, the latter on a 4 point ordinal scale. A section was devoted to satisfaction for different aspects, scores on a 1-10 range. Lastly, planned behaviour, like the intention of looking for another job, was inquired. Optimal scaling analy-

Marisa Civardi, Franca Crippa, Vincenzo Bagnardi
Universita' di Milano-Bicocca e-mail: marisa.civardi@unimib.it, franca.crippa@unimib.it, vincenzo.bagnardi@unimib.it

sis, namely Nonlinear Categorical Principal Component Analysis (CATPCA), was applied to obtain optimal assignment of quantitative values to qualitative data and to reduce dimensionality. Results showed three clearly distinct underling dimensions (Fig.1). The first one concerned objective characteristics of the work: wages, full or part-time, responsibility over other employees, permanent contract or other types of contract, chances of gain in free time and in reconciliation with family bounds. The second dimension, self-realization, played a vital role on its own and it implies job content, like independence as the degree of decisional and organizational autonomy; the extent self fulfillment and expression of one's skills are permitted; future prospects, in terms of promotion and security. The third dimension clearly involved the coherence between graduates' university educational path and the present profession: questionnaire items that constitute this latent variable asked explicitly whether a degree was necessary for the job and, in case of a positive answer, to what extent. Different perspectives were inquired: adequacy of salary and of competencies to the education level, respondence of university training to work requirements.

Table 1 Latent dimensions resulting from the CATPCA analysis

Dimension	Item
Objective features	salary, full or part-time, permanent or temporary contract, free time and reconciliation with family duties
Coherence	necessity of a degree, need for the specific degree, coherence between education and work, adequacy of wage, adequacy of university training to competencies required, future planned behaviour, global satisfaction
Self realization	decisional autonomy, self fulfillment, career prospects, security

An emphasis on only wages and hours of work is likely to give a misleading picture of what makes a good job and hence of graduates' behaviour. Overall measures of job satisfaction should reveal information about rarely measured job content. In the first place, a taxonomy can be build, observing what makes and who has the good jobs. Moreover, determinant of good jobs can be investigated, so as to provide university students with better and better educational paths.

References

Clark A. (2004). *What Makes a Good Job? Evidence from OECD Countries*, Working paper, 2004-28, Paris: OECD.

Fabbris L., Favaro D., Scarsi E. (2010). Un buon lavoro al primo impiego come indicatore di efficacia della formazione universitaria e del capitale umano del laureato. In: (Fabbris L., ed) Dal Bo' all'Agorà. Il capitale umano investito nel lavoro Padova: Cleup, 1-32.

The detection of criticalities of graduates' jobs through Importance-Performance Analysis

Luigi Fabbris, Giovanna Boccuzzo

1 Introduction

Measuring the external effectiveness of the University is a more and more relevant issue, insomuch as the Italian law n.1/2009 indicates to allocate a part of the funds for Universities also on the basis of the quality of educational offering and the output of the academic processes.

For a graduate it is important finding a job as soon as possible after graduation. At the same time, the quality of the job is a crucial aspect, for him/herself and for educational assessment purposes.

In this paper we highlight the critical dimensions of jobs entered by Padua University graduates within six months from graduation. The jobs' critical aspects stem from responses given by graduates in a sample survey carried out by the same Athenaeum for monitoring their employment and professional destinies. The survey involved more than 4800 graduates in 2007 and 2008, the analysis is based on the 2443 employed graduates (Fabbris, 2010).

Jobs were examined from various viewpoints, income, contract, work relationships, worksite, social consideration and personal perspectives. For each job's aspect and for the job as such we measured satisfaction and estimated the importance for Padua graduates.

The technique we adopted for data analysis is IPA - Importance-Performance Analysis (Martilla and James, 1977), a graphical method that can detect criticalities by comparing customers' satisfaction with importance assigned by them to single job aspects. The importance of a job aspect is evaluated by applying a multivariate regression model (Boccuzzo et al., 2010) on various sub-populations, whose importance-performance points are represented on the same graphic. Strata are de-

Luigi Fabbris,
Department of Statistical Science, University of Padua, e-mail: luigi.fabbris@unipd.it

Giovanna Boccuzzo,
Department of Statistical Science, University of Padua, e-mail: giovanna.boccuzzo@unipd.it

fined with reference to type of degree (bachelor versus master), disciplinary area of graduates, whether their work activity started before or after graduation.

2 Main results

The analysis highlighted that the most important factors of job satisfaction are the agreement between job activities and cultural interests and graduates' skill attainment. The lack of coherence between job and university studies and the currently perceived earning perspectives are critical aspects. Career perspectives, work stability, time free of work and specificity of the owned degree get low satisfaction rates but are nevertheless unimportant.

Stratified analyses showed that graduates are heterogeneous: for example, earning perspective is a critical point especially for master graduates, but not for the bachelor ones, probably because people who achieve the higher degree feel greater expectations. Earnings are a secondary issue for graduates at the first job, whereas graduates with a new job are satisfied and graduates that maintain the previous job are less interested in the economic aspects of their job.

Different evaluations are noticeable also by disciplinary area: for graduates in technical fields, unlike others, career's perspectives have satisfaction rates lower than their importance. Maybe they consider that the efforts spent for their degree should be quickly and consistently rewarded by the market.

We draw the conclusion that IPA analysis, especially if it is applied to subgroups of graduates, may help for educational and welfare policy suggestion.

References

Boccuzzo G., Fabbris L., Scarsi E. (2010). Non tutto l'oro luccica. Criticità dei lavori dei laureati identificate tramite Importance-Performance Analysis. In: Fabbris L. (eds.) Dal Bo' all'Agorà. Il capitale umano investito nel lavoro, Cleup, Padova: 95-138.

Fabbris L. (2010). Il Progetto Agorà dell'Università di Padova. In: Fabbris L. (a cura di) Dal Bo' all'Agorà. Il capitale umano investito nel lavoro, Cleup, Padova: V-XLVI.

Martilla J., James J. (1977). Importance-Performance Analysis. *Journal of Marketing*, 41 (1), 77-79.

Statistical methods to describe professional profiles through competences and activities

Cristiana Martini

1 Introduction

The labour market is a mutable reality, where professional profiles change, and new professional roles emerge and/or replace the old ones. This is especially true when we consider innovative services or developing sectors, and even more for the highly qualified jobs; in this case, the new profiles are often unclear also to the agencies that are called to educate people for these professional roles. An effective way to depict a work role is by means of the competences (knowledge, skills and attitudes) which are required to cover that position; different mixtures of competences give rise to different professional profiles, which can be described by analysing a competence-by-job rectangular $(C \times J)$ matrix, whose (c, j)-th element represents the importance of the c-th competence for the j-th job. As an alternative, a job can be described by means of the work activities that are performed, analysing an analogous activity-by-job rectangular $(A \times J)$ matrix, whose (a, j)-th element represents the presence of the a-th activity in the j-th job.

Several multivariate techniques can be used to analyse and graphically represent matrices like these, which focus on different aspects of the mutual relationships between professional profiles, competences and activities. Aim of this contribution is to show how these techniques can be used to describe professional profiles, and what kind of complementary information they provide.

2 Main results

The analyses are focused on the professional profiles operating in the Research and Development (R&D) field. Professional roles and competences were surveyed by

Cristiana Martini,
University of Modena and Reggio Emilia, e-mail: cmartini@unimore.it

interviewing the directors of a sample of 31 (out of 66) R&D companies with at least 3 operators in the Veneto Region. The face-to-face interviews focused on the professional profiles employed in the firm (at least at a technical level); for each profile, directors were asked how many people in such position were employed in the firm, what activities they had to perform in their working tasks, and what competencies were needed to cover each position. The final dataset contains professional profile, educational level, performed activities and required competences of 450 persons who work in the R&D field.

Correspondence analysis conducted on activities allows to generate a graphical representation of the two main dimensions: one runs from technical to managerial activities, the other from operative to scientific tasks. Professional profiles and competences can then be projected on these dimensions as supplementary points, in order to describe each job in terms of involved activities and required competences, and to identify similar professional profiles. The analysis underlines the existence of 3 main professional areas: a managerial area, a scientific area, and a technical-operational area. In principle, an analogous operation can be performed on the basis of performances, but competences are in general more cross-occupational, and their capability to describe professional profiles is lower.

The interconnection between roles and competences can also be analysed by means of a network approach; the data, in fact, can be seen as a 2-mode affiliation network (job-competence, or alternatively job-activity). The 2-mode network is then transformed according to a 1-mode approach, giving rise to separate 1-mode networks for activities, competences, and jobs. In particular, the competence network will connect those competences which tend to be jointly required in the same jobs; analogously, the ability network will connect abilities that are performed by the same professional profiles; finally, two alternative job networks are possible: one connecting profiles which require the same competences, the other linking jobs characterised by common activities.

The 1-mode job network based on competences (where the connections indicate common required competences), shows a very strong interconnection between the core profiles of the R&D firm: director, scientific manager and director. Strong connections are observed also between this core group and the technical manager or the laboratory technician; connections are relevant with samplers, commercial managers, coordinators and executors too, while the other profiles show a lower similarity, and some are completely set apart. The 1-mode job network based on activities, with connections indicating common activities across roles, is a much less connected network, indicating that different roles may require the same competences, but the tasks distribution is quite clear. These network structures hint a clear system of professional relationships among professional profiles, which in some sense reproduces a sort of operational organogram.

In conclusion, describing professional profiles by means of competence-by-job and activity-by-job matrices gives an in-depht description of each professional role, and allows to depict the professional profiles of an occupational sector, and their mutual relationships.

Contributed Session 21

Correspondence Analysis and Related Methods I

Joint Correspondence Analysis vs. Multiple Correspondence Analysis: a solution to an undetected problem

Sergio Camiz, Gastão Coelho Gomes

1 Introduction

Multiple Correspondence Analysis (*MCA*) is the most evident result of the extension of Principal Component Analysis to other kind of data, a common result of several points of view, developed independently but with the same aim: the detection of common quantitative independent factors that may act, simultaneously and at the best, as the interpretation of the relations among several nominal characters and graphical tools for their common representation.

Despite the broad interest and large use by the data analysts, its application is not appreciated by statisticians that consider a drawback its reduced amount of inertia attributed to the first factors. This may be one of the reasons that drew both Benzécri (1979) and Greenacre (1988) to re-evaluate it according to two different criteria. Whereas Benzécri grounds his theoretical arguments on character levels resulting by discretization of continuous normally distributed variables, Greenacre puts in evidence that *MCA* is too heavily influenced by the Burt's matrix block diagonal sub-matrices. Indeed, *MCA*, as Single Correspondence Analysis (*SCA*) of the Burt's table, should partition the deviation from expectation into principal components, so that the block diagonal sub-matrices have no interest. Thus, he developed the Joint Correspondence Analysis (*JCA*), in order to overcome the problem and at the same time to reevaluate the inertia explained by the factors whose corresponding eigenvalue is larger than their mean value.

At a first glance, the results of *JCA*, in comparison with the results of *MCA*, do not differ so much, I would say not enough to convince a user to convert to *JCA*, that may be run with *R* through ca procedure, shifting from some well established *MCA* program able to provide full interpretation aids and wonderful graphics. The

Sergio Camiz,
Sapienza Università di Roma, e-mail: sergio.camiz@uniroma1.it

Gastão Coelho Gomes,
Universidade Federal do Rio de Janeiro, e-mail: gastao@im.ufrj.br

things change, when the reconstruction of the Burt's table is inspected step by step. The idea was applied by Orlóci (1978) to *SCA*, but no trace results so far in *MCA* studies. I shall use this point of view to compare the two methods.

2 Main results

When applied to a single contingency table, a relation between the eigenelements of *SCA* and *MCA* on both indicators and Burt's matrix is known: given λ_α, μ_α, and ν_α the respective eigenvalues, sorted in descending order, it results that $\nu_\alpha = {\mu_\alpha}^2$ and $\nu_\alpha = \frac{1 \pm \sqrt{\lambda_\alpha}}{2}$, with the minus sign that explains the high number of small, unuseful eigenvalues. According to both authors they may be dropped and, generalizing, the same argument is applied to tables with more characters.

Unfortunately, in the description of his new proposal, Greenacre (1988) does not show in practice an important point in favour of *JCA*, that is its better reconstruction of the off-diagonal submatrices of the Burt's table, an important plus in favour of *JCA*. Indeed once the reconstruction formula is applied to the off-diagonal submatrices, the performance of *JCA* is not only better in respect to *MCA* equal dimensions solution, but it is monotone whereas the other, unlike the common belief, is not. In an example, taken from Camiz and Coelho Gomes (2009), this pattern will be shown, in which the *MCA* reconstruction of the off-diagonal elements becomes worst and worst, until eventually adjusted thanks to the last eigenelements. Thus, the drop of the least eigenelements loses its rationale, with important consequences.

It is strange that nobody until now put in evidence this point, that is by no means a very serious drawback of *MCA*. On the opposite, considering the *MCA* rationale, the interpretation of this result is very easy. As *MCA* is just a *SCA* of the Burt's table and since *SCA* extracts first the highest deviations from the expectation, what could be a larger deviation unless the diagonal submatrices themselves, with their maxima on the diagonal and zeros elsewhere?

References

Benzécri, J. P. (1979), "Sur les calcul des taux d'inertie dans l'analyse d'un questionnaire", *Les Cahiers de l'Analyse des Données*, 4(3): pp. 377-379.

Camiz S., Coelho Gomes G. (2009), "Correspondence Analyses for Studying the Language Complexity of Texts", VIII Congreso Chileno de Investigación Operativa, OPTIMA, Concepción (Chile), on CD-ROM.

Greenacre M. (1988), Correspondence analysis of mutlivariate categorical data by weighted least squares, *Biometrika*, 75: pp. 457-467.

Orlóci L. (1978), *Multivariate Analysis in Vegetation Research*, 2nd ed., den Haag, Junk.

Measuring subcompositional incoherence in contingency tables and compositional data

Michael Greenacre

1 Summary

Subcompositional coherence is a fundamental property of Aitchison's log-ratio approach to compositional data analysis, and is the principal justification for using ratios of components (Aitchison, 1983). We maintain, however, that lack of subcompositional coherence, that is incoherence, can be measured in an attempt to evaluate whether any given technique is close enough, for all practical purposes, to being subcompositionally coherent (Greenacre, 2008). This opens up the field to alternative incoherent methods such as correspondence analysis (CA), which might be better suited to cope with problems such as data zeros and outliers, while being only slightly incoherent. The same idea can be applied to the analysis of contingency tables, to which both log-ratio analysis and CA can be applied. The measure that we propose is based on the distance measure between components. We show that the two-part subcompositions, which are the most sensitive to subcompositional incoherence, can be used to establish a distance matrix which can be directly compared with the pairwise distances in the full composition. The closeness of these two matrices can be quantified using a stress measure that is common in multidimensional scaling, providing a measure of subcompositional incoherence. Furthermore, we strongly advocate introducing weights into this measure, where rarer counts or components are weighted proportionally less than more frequent counts or components (Greenacre and Lewi, 2009). It is shown how power-transformed CA (which converges to log-ratio analysis as the power parameter tends to 0 – Greenacre, 2009, 2010) can be used to give the least incoherent CA as an alternative to log-ratio analysis when the data contain many zero values.

Michael Greenacre,
Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, Barcelona, Catalunya, Espanya.
e-mail: michael@upf.es

References

Aitchison J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65.

Greenacre M., Lewi P. J. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio scale measurements. *Journal of Classification* 26, 29–54.

Greenacre M. (2008). Measuring subcompositional incoherence. Working paper n.1106, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona. URL:http://www.econ.upf.edu/en/research/onepaper.php?id=1106. Provisionally accepted for publication in *Mathematical Geosciences*.

Greenacre M. (2009). Power transformations in correspondence analysis. *Computational Statistics and Data Analysis* 53, 3107–3116.

Greenacre M. (2010). Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences* 42, 129–134.

Correspondence Analysis in the case of outliers

Anna Langovaya, Hamdi Chouikha, Sonja Kuhnt

1 Introduction

In correspondence analysis as well as in every statistical analysis, observations can appear that seem to deviate strongly from the majority of the data. Such observations are usually called outliers and may contain important information about unknown irregularities, dependencies and interactions within the data.

As concerns correspondence analysis, outliers are given by specific cell frequencies of the underlying contingency table. Situations can occur where outliers are present in the table which are not immediately suspicious but play a crucial role for the statistical analysis. In such cases our approach will be useful.

2 Main results

In our research we apply correspondence analysis (Blasius and Greenacre, 2006) to three-way contingency tables with data entries generated by deviations from the independence model. Specific dependencies are caused by outliers of moderate size. A formal definition of outliers in the context of contingency tables is given in Kuhnt (2004).

Outliers in our work are chosen in such a way that they break the independence in the table, but yet they are not large enough to be easily spotted without statistical analysis. We study the change in the correspondence analysis row and column co-

Anna Langovaya,
TU Dortmund University, e-mail: langovaya@statistik.tu-dortmund.de

Hamdi Chouikha,
TU Dortmund University, e-mail: chouikha@statistik.uni-dortmund.de

Sonja Kuhnt,
TU Dortmund University, e-mail: kuhnt@statistik.tu-dortmund.de

ordinates caused by the outliers. We also perform numerical analysis of the outlier coordinates and suggest possible criteria for identifying hidden outliers in multiway contingency tables by means of correspondence analysis coordinates.

References

Blasius J., Greenacre M. (2006). *Multiple Correspondence Analysis and Related Methods,* London: Chapman & Hall/CRC.

Kuhnt S. (2004). Outlier Identification Procedures for Contingency Tables using Maximum Likelihood and L_1 Estimates. *Scandinavian Journal of Statistics,* 31, 431-442.

Contributed Session 22

Estimation Problems

Principal Stratification in sample selection problems with non normal error terms

Giovanni Mellace, Roberto Rocci

1 Introduction

Post-treatment complications such as non-ignorable non response may affect inference on causal effects in both observational and experimental studies. Frangakis and Rubin (2002) have proposed principal stratification to deal whit these kind of problems. In this approach point identification is achieved by means of a finite mixture of Gaussian distributions (McLachlan and Peel, 2000).

In this paper we propose a principal stratification mixture of mixture estimator, which can effectively deal with missing outcome problems, in presence of non normal error terms.

In order to study the performance of our estimator, we run a Monte Carlo simulation in which we compare the performance of our estimator and the Heckman maximum likelihood and two-step estimators (Heckman, 1974). The results show that our approach performs better in terms of mean square error (MSE) when the true distribution of the error terms is far from being Gaussian.

2 Main results

Suppose we want to estimate the effect of a binary treatment $t = 1,0$, on an outcome y, at a specific time after assignment. We will denote by $Y_i(1)$ and $Y_i(0)$ the two potential outcomes according to the SUTVA (Stable Unit Treatment Value As-

Giovanni Mellace,
Università di Tor Vergata, via Columbia 2, 00133 Roma, Italy,
e-mail: Giovanni.Mellace@uniroma2.it

Roberto Rocci,
Università di Tor Vergata, via Columbia 2, 00133 Roma, Italy,
e-mail: Roberto.Rocci@uniroma2.it

sumption). Let $q_i = 1,0$ be equal to 1 if we observe the outcome of individual i and 0 otherwise. We assume that the treatment is unconfounded given a vector x_i of J pre-tretment variables.

Principal stratification suggests to stratify the units, within each cell defined by the values of the covariates, into four latent principal strata, according to the joint values of $(q_i(1), q_i(0))$ that we denote by $\{11, 10, 01, 00\}$. We assume monotonicity so that no units belong to 01.

Then, the likelihood function results in a finite mixture of distributions, where the population proportions of units belonging to each stratum in the cell $X = x$, denoted by $\pi_{11|x}$, $\pi_{10|x}$, and $\pi_{00|x} = 1 - \pi_{11|x} - \pi_{10|x}$, are modeled as a multinomial logit. The distributions of y conditional to the principal strata are often assumed to be

$$f(y_i|11, x_i, q_i = 1) \sim \mathcal{N}(\beta_0 + \beta_1 t_i + \beta_2^T x_i, \sigma^2)$$
$$f(y_i|10, x_i, q_i = 1) \sim \mathcal{N}(\delta_0 + \delta_1 t_i + \beta_2^T x_i, \sigma^2)$$

where we set $\delta = \delta_0 + \delta_1 t_i$ since t_i is always equal to 1 in stratum 10.

It follows that conditionally on stratum 11, we have that $y_i = \beta_0 + \beta_1 t_i + \beta_2^T x_i + \varepsilon_i$, while conditional on stratum 10, $y_i = \delta + \beta_2^T x_i + \varepsilon_i$. The intuition of our approach is the following. Maintaining all the above assumptions, we assume $f(\varepsilon_i) = \sum_{g=1}^{G} \tau_g \phi(\varepsilon_i; \mu_g, \sigma^2)$,i.e. we approximate the true error distribution by a mixture of Gaussians. So that we have

$$f(y_i|11, x_i) = \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \beta_0 + \beta_1 t_i + \beta_2^T x_i, \sigma^2)$$

$$f(y_i|10, x_i) = \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \delta + \beta_2^T x_i, \sigma^2).$$

where $\tau_g \geq 0,\ g = 1, \ldots, G, \sum_{g=1}^{G} \tau_g = 1, \sum_{g=1}^{G} \tau_g \mu_g = 0$ and $\phi(y_i, \mu, \sigma^2)$ denotes the density at y_i of the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

To check the effectiveness of our estimator, we implemented a Monte Carlo simulation, where samples have been generated according to a standard Heckman sample selection model with five different non-normal distributions for the error terms. The results show that our estimator performs better in terms of MSE, than the Heckman maximum likelihood and two-step estimators, with all the distributions considered.

References

Frangakis C. E., Rubin D. B., (2002). Principal stratification in causal inference. *Biometrics* 58, 191–199.

Heckman J., (1974). Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–694.

McLachlan G., Peel D., (2000). *Finite Mixture Models*, New York: Wiley series in probability and statistics.

Some considerations on two-stage calibration estimators

Mario Montinaro, Ivan Sciascia

1 Introduction

In this paper we introduce a calibration estimator suitable for sparse spatial data. In particular the sampling and estimation procedures can be applied to spatial data analysis where certain variables are remotely sensed. We propose to correct to the calibration estimators by means of a function that takes into account the total distance between two sampled units.

2 Main results

A previous study of Opsomer et al. (2007) considered the application of estimators for spatial data of forest resources. We define a grid of sparse data where the units meet the following space constraint:

$$\sharp(z_i) \leq 2 \tag{1}$$

where $\sharp(z_i)$ is the number of units included in the minimum space unit z_i. Consider a finite population of N units where associated with the ith unit are the study variable y_i and the auxiliary variable x_i. If we consider c_i as calibrated weights, w_i the Horvitz-Thompson weights and $G_i(c_i, w_i)$ the chi-squared distance function the optimization problem that leads to the calibration estimator is:

Mario Montinaro,
Dipartimento di Statistica e Matematica applicata, Universita' di Torino,
e-mail: mario.montinaro@unito.it

Ivan Sciascia,
Dipartimento di Statistica e Matematica applicata, Universita' di Torino
e-mail: ivan.sciascia@unito.it

$$c_i : min \sum_{i=1}^{n} G_i(c_i, w_i) \tag{2}$$

$$\text{subject to} \sum_{i=1}^{n} c_i x_i = X \tag{3}$$

that, in turns, yields the calibration estimator of Deville and Sarndal (1992):

$$\hat{Y}_c = \sum_{i=1}^{n} c_i y_i = \hat{Y} + \hat{\beta}_c (X - \hat{X}) \tag{4}$$

$$\hat{\beta}_c = \sum_{i=1}^{n} w_i x_i y_i / \sum_{i=1}^{n} w_i x_i^2 \tag{5}$$

Our correction to $\hat{Y}_c$ involves the distance between sampled units in the minimum space unit and its nearest neighbor in the following way:

$$\hat{Y}_{cs} = \hat{Y}_c + f(d_T) \tag{6}$$

where $f(d_T)$ is a function of distance between the sampled unit and its nearest neighbor.

Note that, if the distance is linearly approximated, then it can be computed by using the coordinates of the units. For (x_1, y_1), (x_2, y_2) two sampled units, the distance is calculated as:

$$d_T = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} * 2 \tag{7}$$

Further work will focus on the explanation of the distance function in the construction of the estimator proposed.

References

Deville J. C., Sarndal C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376–382.

Opsomer J. D., Breidt F. J., Moisen G. G., Kauermann G. (2007). Model-Assisted Estimation of Forest Resources With Generalized Additive Models. *Journal of the American Statistical Association*, 102 (478), 400-409.

Automatic strategies of analysis for handling structurally missing occasions

Nadia Solaro

1 Introduction

In most fields of application, data are of three-way type. This intends that, in addiction to objects and variables, data contain a third dimension, typically represented by time or, more in general, data collection occasions. From a statistical point of view, three-way data can be processed in explorative terms through statistical methods known in the literature as multiway data analysis techniques (Kroonenberg, 2008).

A problem that can arise with three-way data is that the number of collection occasions can differ over objects. This situation arises in particular whenever the occasion number depends on whether objects attain or not a predetermined set of conditions. In a previous work, we denoted this situation with the term of *structurally missing occasions* (Solaro and Vittadini, 2010). This concept can be extended to that of pseudo-structurally missing occasions, to comprise more general situations in which objects cannot be exposed to the same number of data collections.

The main consequence of this missingness pattern is that multiway data analysis techniques can no more be applied in their standard form. They would require in fact that all objects have values on all variables at all occasions (Kroonenberg, 2008). To our knowledge, the issue of a different number of occasions over objects has not been yet systematically tackled in the literature, since most contributions in this area are devoted to imputation techniques (see, e.g., Kroonenberg, 2008, Chap. 7). To overcome the problem of (pseudo-)structurally missing occasions, in this work we propose a range of strategies aimed at completing data matrix rows pertaining to objects that from a certain occasion onward exhibit (pseudo-)structurally missing data.

Next, three-way stress multidimensional scaling (Borg and Groenen, 2005) will be applied to produce low-dimensional configurations of points pertaining to each occasion as well as a common space. Finally, the extracted dimensions will be monitored along the occasions through trajectory plots, with the primary concern of highlighting different information contents inherent in the various strategies.

Nadia Solaro, Department of Statistics, University of Milano-Bicocca, Italy
e-mail: nadia.solaro@unimib.it

2 Main results

Assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ is a sequence of data matrices of dimensions $n_k \times p$ collected on K occasions, with $n_1 = n$ being the object set size at the first occasion and $n_k \leq n_{k-1} \leq n$ the object set size at subsequent occasions, where the strict inequality holds for at least one k, $(k = 2, \ldots, K)$. Then, to have data matrices of the same dimension we propose to augment each matrix $\mathbf{Y}_k$ by assigning suitable values to the $(n - n_k)$ objects that are absent at the k-th occasion.

In a previous work, we made a first proposal for row augmentation by carrying forward the most recent observations available on these objects (Solaro and Vittadini, 2010). We termed this strategy "LOCF-row augmentation" (LOCF, Last Observation Carried Forward), which is drawn from the well-known idea of LOCF imputation of missing data (Molenberghs and Kenward, 2007). Since in general variables require to be standardized before proceeding to analyses, the LOCF-row augmentation approach can then result in two different strategies: (1) FIAMS (FIrst Augment the Matrix, then Standardize), in which standardization is carried out only after having augmented matrix $\mathbf{Y}_k$ with LOCF-values; (2) FISAM (FIrst Standardize, then Augment the Matrix), in which matrix $\mathbf{Y}_k$ is firstly standardized and then augmented with standardized scores carried forward from each object's last available standardization.

Next to these, two other strategies will be introduced and then compared in the present work. Specifically, these are: (3) zero-row augmentation strategy, which consists of row-augmenting each matrix $\mathbf{Y}_k$, previously standardized, by adding zero values for the objects absent at the k-th occasion. In this way, they would represent a sort of reference category, in that their values would coincide with the average of standardized variables; (4) zero-distance augmentation strategy, which, unlike the other strategies, consists of augmenting each dissimilarity matrix $\boldsymbol{\Delta}_k$ by adding zero values for the objects absent at the k-th occasion. In this way, absent objects would be treated as if they were coincident from the point of view of the information status.

References

Borg I., Groenen P.J.F. (2005). *Modern Multidimensional Scaling*, 2nd edition, New York: Springer.

Kroonenberg P.M. (2008). *Applied Multiway Data Analysis*, New Jersey: John Wiley & Sons.

Molenberghs G., Kenward M.G. (2007). *Missing Data in Clinical Studies*, Chichester: John Wiley & Sons.

Solaro N., Vittadini G. (2010). Assessing individual treatment effectiveness in the presence of structurally missing measurement occasions. *to appear*.

Contributed Session 23

Issues in Classification and Clustering

Model-based clustering of multistate data with latent change. An application with DHS data

José G. Dias

Finite mixture modelling has been used extensively as a model-based clustering technique. This research addresses the application of mixture models to multistate data (sequences of states) under the Markov assumption. By assuming a latent or hidden Markov process, the model allows for misclassification error (Baum et al., 1970). The data are from the life history calendar from the Brazilian Demographic and Health Survey 1996 (BENFAM and Macro International, 1997) in which contraceptive use dynamics are surveyed retrospectively. The results show that the dynamics are heterogeneous with three subpopulations. Moreover, this research extends results reported in Dias and Willekens (2005), in which contraceptive use dynamics are modelled as a manifest process rather than a latent one.

References

Baum L.E., Petrie T., Soules G., Weiss N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41, 164-171.

BENFAM and Macro International (1997). *Pesquisa Nacional Sobre Demografia e Saúde (PNDS), Brasil, 1996*, Rio de Janeiro: Brasil.

Dias J.G., Willekens F. (2005). Model-based Clustering of Sequential Data with an Application to Contraceptive Use Dynamics. *Mathematical Population Studies*, 12(3), 135-157.

José G. Dias,
Department of Quantitative Methods & UNIDE, ISCTE – Lisbon University Institute,
e-mail: jose.dias@iscte.pt

Issues on clustering and data gridding

Jukka Heikkonen, Domenico Perrotta, Marco Riani, Francesca Torti

1 Introduction

In this paper we address a clustering issue in a situation where the data consist in an unknown number of groups with high degree of overlapping. The dataset in the left panel of Figure 1, coming from international trade data, well exemplifies this situation. This dataset has been treated in Perrotta and Torti (2009) as multivariate problem with the Forward Search (FS). The FS is a method by Atkinson et al. (2004) for detecting unidentified subsets and masked outliers and for determining their effect on models fitted to the data. In the random starts version of the FS it is possible to monitor the trajectories of the values of a particular statistic, e.g. the minimum Mahalanobis distance among observations outside the subset. An example of this technique is shown in the right panel of Figure 1. In general the presence of multiple populations is revealed as separated peaks in the minimum Mahalanobis distance plot. However, problems may occur in presence of high density areas. For example in the left panel of Figure 1 more than 50% of the data are concentrated near the origin of the axes and, thus, the trajectories of Minimum Mahalanobis distance degenerate into the same search path in the very first steps of the FS. More in general, similar difficulties arise with other classical methods such as Kmeans clustering, mixtures of Gaussians, etc.

In this work we present a possible approach to avoid the disturbing effects of the dense populated data points through a data gridding technique based on Principal Component Analysis (PCA). This technique consists in defining a grid along the Principal Component (PC) axes of the data and selecting one point in each cell of the grid.

Jukka Heikkonen and Domenico Perrotta,
European Commission, Joint Research Centre, Ispra, Italy,
e-mail: jukka.heikkonen@jrc.ec.europa.eu – domenico.perrotta@ec.europa.eu

Marco Riani and Francesca Torti
University of Parma, Italy, e-mail: mriani@unipr.it – francesca.torti@nemo.unipr.it

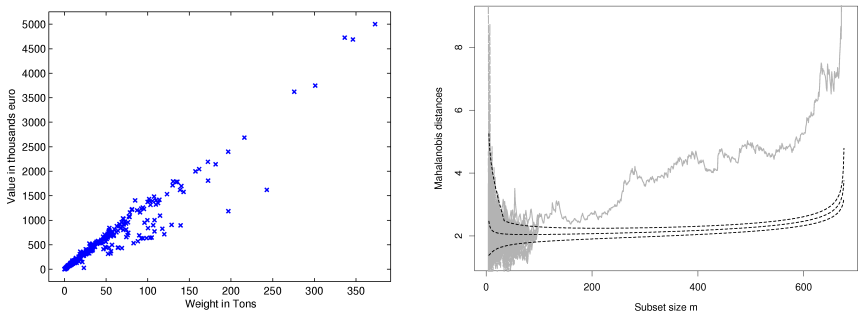


Fig. 1 An example of data with at least two populations and random starts Minimum Mahalanobis Distance plot with 1%, 50% and 99% confidence bands under the null hypothesis of a single normal population.

2 Main results

Taking the maximum and minimum coordinates of the PCs we can define a grid of predefined number of cells (e.g., 50) along PC axes. In our case when we have 2-dimensional data, the grid is also 2-dimensional and defined by 2 PC axes. The same approach can be extended to higher dimensions. Note that the cells do not necessarily have equal width in all PC axes directions and a single cell can cover zero, one or more data points. Especially where the original data is densely populated, the corresponding grid cell includes multiple samples. For each cell the goal is to search one representative sample from the original data; this is performed by taking the median of samples belonging to the cell and finding the closest sample to the calculated median. As a result we have either none or 1 sample for each cell of the grid. Then with the new reduced subset we can perform a desired analysis, for example estimating the mixture model. Finally, all observations in the original dataset can be assigned to the estimated mixture according to simple distance criteria.

The paper will demonstrate the effectiveness of this approach with FS and Gaussian mixture clustering models.

References

Atkinson A.C., Riani M., Cerioli A. (2004). *Exploring Multivariate Data with the Forward Search*, New York: Springer.

Perrotta D., Torti F. (2009). Detecting price outliers in European trade data with the forward search, in Data analysis and classification: from exploration to confirmation, Springer studies in classification, data analysis, and knowledge organization, pp 415-423.

Tree partitioning criteria across objects and predictors for data with a double stratification

Valerio A. Tutore, Valentina Cozza and Antonio D'Ambrosio

1 Introduction

Understanding complex data structures in large databases is the new challenge for statisticians working in various fields such as finance, biology, and so on. This complexity often refers to the dimensionality of the units or the variables, or it is referred to both of them.

The framework of this work is supervised learning using classification and regression trees (Breiman *et al.*, 1984). These models explore the relations between a set of predictors and a response variable generating a set of rules that support the decision making processes in the most different situations. As a matter of fact, dealing with complex relations among the variables, every CART-based approach offers unstable and not interpretable solutions.

This work has been inspired by an analysis of a particular data set coming from a survey collected at several Italian research institute. Our real problem was the complex data structure: data were indeed characterized by a multiple instrumental variable, whose role is the stratification of both the statistical units and the variables. These data can generally be summarized as follow:

let Y be the response variable which can be either categorical or numerical, and let $\mathbf{X} = \{X_1, \ldots, X_M\}$ be the set of M predictors, or inputs. In the following we consider the role played by instrumental variables to stratify both the predictors and the

Valerio A. Tutore,
Dipartimento di Matematica e Statistica, Universitá di Napoli Federico II.
e-mail: v.tutore@unina.it

Valentina Cozza,
Dipartimento di Studi Aziendali, Universitá di Napoli Parthenope.
e-mail: valentina.cozza@uniparthenope.it

Antonio D'Ambrosio,
Dipartimento di Matematica e Statistica, Universitá di Napoli Federico II.
e-mail: antdambr@unina.it

objects. In other words, we deal with the case in which the input variables are stratified into G groups on the basis of an instrumental variable Z_g and, simultaneously, the objects are stratified into K categories on the basis of a stratifying variable Z_k.

These situations have been separately handled in the recent past for classification (Tutore *et al.*, 2007; Tutore and D'Ambrosio, 2009) and regression problems (Giordano and Aria, 2010).

In this work we propose a combination of the Two-Stage segmentation via DIScriminant analysis (TS-DIS) (Mola and Siciliano, 2002) and the conditional splitting criteria for three-way data matrix as defined in Tutore *et al.* (2007) or in Tutore and D'Ambrosio (2009).

The TS-DIS algorithm helps in reducing the dimensionality of the analysis, moving the focus to a set of new latent predictors, that summarize the original variables. This characteristic makes the method very convenient when treating big sample of data shaped according with a stratification variable.

The conditional splitting criteria are based on conditional impurity measures which are a weighted average of each partial impurity in every class of the Z_k stratifying variable.

The combination of both the methodologies into a new recursive partitioning algorithm allows to obtain a new tree structure that combines their single interpretative improvements without losing in accuracy, as it is confirmed by several analysis on both simulated and the previously described real-world data sets.

References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Giordano, G., Aria, M. (2010) Regression trees with moderating effects. *Springer series in New Perspectives in Statistical Modeling and Analysis*, forthcoming.

Mola, F., Siciliano, R. (2002). Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining, in Roli, F., Kittler, J. (eds.): *Proceedings of International Conference on Multiple Classifier Systems* (Chia, June 24-26, 2002), 118-126, Lecture Notes in Computer Science, Springer, Heidelberg.

Tutore, V.A., Aria, M., Siciliano, R. (2007). Conditional classification trees using instrumental variables. *LNCS 4723, Advances in Intelligent Data Analysis*, Springer-Verlag, pp 163-173.

Tutore V.A., D'Ambrosio A. (2009).Three-Way Data Analysis by Tree-Based Partitioning. *Classification and Data Analisys 2009, Book of short papers* (Catania, September 9-11, 2009), CLEUP Padova, 641-644.

Contributed Session 24

Multivariate Analysis for Relational Data

Cluster analysis of multivariate relational data

Vladimir Batagelj

1 Introduction

The *multivariate relational data* can be described by a network $N = (V, L, P, w)$, where V is the set of *vertices* (units), L is the set of *lines* (arcs or edges), $P = \{p_1, p_2, \ldots, p_s\}$ is the set of *vertex properties* (variables), and w is the *weight* on lines L. The vertex properties determine a multivariate table $X = [x_{vi}]_{V \times P}$. If $w \equiv 1$ we call the data the *simple multivariate relational data*.

A general question is how to extend or reformulate classical multivariate analysis problems to account also for the additional relational data. In this paper we present some ideas how to approach the clustering problem.

2 Main results

There are several clustering based approaches to analyze the multivariate relational data.

Network analysis of multivariate data. We start with the multivariate table X and selected dissimilarity d between its rows (units). The set of lines L is determined from the data either (a) as all pairs of vertices $(u : v)$ (edges) for which $d(u,v) \leq r$, where $r > 0$ is a selected radius; or (b) as all ordered pairs of vertices (u,v) (arcs) for which v is among the k closest neighbors of u, where $k > 0$ is a selected integer. The weight of a line (u,v) equals to the dissimilarity between its endpoints $w(u,v) = d(u,v)$.

Vladimir Batagelj,
University of Ljubljana, Faculty of Mathematics and Physics, e-mail: vladimir.batagelj@fmf.uni-lj.si

On the obtained network (V, L, w) we can use different network analysis methods (Wasserman and Faust, 1994), such as connected components, strong components, (generalized) cores, islands, etc., to identify clusters of similar vertices (units).

Clustering with relational constraints and blockmodeling. In the case of simple multivariate relational data we can use the multivariate data to determine the clustering criterion function and relational data to determine the set of feasible clusterings. In clustering with relational constraint each cluster in feasible clustering should induce a subgraph of selected type of connectivity (Ferligoj and Batagelj, 1983; Gordon, 1996). In blockmodeling each block $L \cap C_i \times C_j$, where C_i and C_j are clusters, should be as close as possible to selected types of blocks (Doreian et al., 2005).

Recently we developed a very efficient algorithm for hierachical clustering with relational constraints of large sparse networks.

Extending multivariate data with structural properties. An approach to reduce the problem to standard multivariate analysis is to add to the set of vertex properties P some additional structural properties (variables) computed from the network structure, such as (Wasserman, 1994): degree, indegree, outdegree, betweenness, clustering coefficient, core number, etc. The obtained extended multivariate table is afterward analyzed using standard multivariate analysis methods.

Optimization clustering problems. On the basis of multivariate relational data different (multicriteria (Ferligoj and Batagelj, 1992)) optimization clustering problems can be formulated following the direct or indirect approach described in details in (Batagelj and Ferligoj, 2000). This is almost untouched topic for research – only Aleš Žiberna did some work on blockmodeling of valued networks.

References

Batagelj V., Ferligoj A. (2000). In: Clustering relational data. Data Analysis (ed.: W. Gaul, O. Opitz, M. Schader), Berlin: Springer, 3–15.

Doreian P., Batagelj V., Ferligoj A. (2005). *Generalized Blockmodeling*, Cambridge: Cambridge University Press.

Ferligoj A., Batagelj V. (1983). Some types of clustering with relational constraints. *Psychometrika* 48, 4, 541–552.

Ferligoj A., Batagelj V. (1992). Direct Multicriteria Clustering Algorithms. *Journal of Classification* 9, 1, 43–61.

Gordon A.D. (1996). A survey of constrained classification. *Computational Statistics and Data Analysis* 21, 17–29.

Wasserman S., Faust K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Multiple Correspondence Analysis for relational data

Domenico De Stefano and Giancarlo Ragozini

1 Introduction

Correspondence analysis (CA) has been frequently used in social network analysis (SNA), to analyze and graphically represent two-mode networks (Wasserman et al., 1989; Roberts, 2000). According to Wasserman et al. (1989), CA is a data analytic technique used to study contingency tables which is applicable in a variety of ways to relational data. Its use is principally due to the similarity between the affiliation matrix **F** associated to a two-mode network and the usual contingency table. However, as CA is not designed to treat relational data its use in social network analysis has been criticized by some authors (Borgatti and Everett, 1997). Furthermore, in our opinion, CA is not appropriate and correct approach in dealing with binary affiliation matrices even if it could lead to meaningful results.

Hence, aiming at exploring and visualizing the relational structure of a two-mode network, we propose a more suitable approach based on an appropriate transformation of the raw affiliation matrix. In particular, we propose the use of multiple correspondence analysis (MCA) for two-mode networks. MCA has been used by some authors with different purposes (Wasserman et al., 1989). Here we follow a different approach which consists in CA performed on a special case of indicator matrix **Z** build up from **F** and in the use of the concept of "doubling" (Greenacre, 1984). We will show that MCA, thanks to its properties, presents some notable advantages with respect to CA.

Domenico De Stefano
Dept. of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste,
e-mail: domenico.destefano@econ.units.it

Giancarlo Ragozini
Dept. of Sociology, University of Naples Federico II, Vico Monte della Pietá - Napoli,
e-mail: giragoz@unina.it

2 Multiple correspondence analysis for two-mode networks

Formally, a two-mode network is an object $G(V_1, V_2, R)$ consisting of two disjoint sets respectively of I actors and J events. $\mathbf{F}$ is a (binary) $I \times J$ actor-by-relation matrix. The use of CA for such data is fully justified only if it is assumed that the affiliation matrix is frequency-like (Roberts, 2000). However, such assumption cannot be generalized to every relational dataset, especially for binary data. Hence, CA could fail to reconstruct the relational structure embedded in the data.

Focusing the analysis on the actors, we propose to consider the relational events as "bipolar variables": the participation to a given event e_j is treated as a 2-point scale whose positive and negative poles consist respectively in actor participation (e_j^+) and actor non-participation (e_j^-). In such a way we construct an $I \times 2J$ "doubled" indicator matrix $\mathbf{Z}$ from the original matrix $\mathbf{F}$, whose rows are indexed by the actors and the columns by "bipolar" event variables (see Greenacre, 1984 for details on doubling). MCA is performed by applying a standard CA on $\mathbf{Z}$. We will show that this approach leads to a "better" representation of the actor subgroups in which distances among them could be directly interpreted as a measure of the strength of subgroup affiliation. Indeed, with respect to CA, the peculiar properties of MCA allow that the actor coordinates express just their participation patterns, and not also other features like the actor degrees or the number of participants to a certain event, as in classical CA. The asymmetric treatment of rows and columns leads to several advantages with respect to CA in both actor and event subspaces. In the actor subspace the interpretation is improved thanks to the following properties: *i*) actors are close to each other only because of similarities in their participation patterns, independently on their degrees; *ii*) total inertia depends solely on the differences in actor relational patterns; *iii*) weighted euclidean distance among actors are nicely interpreted as a (weighted) simple matching; *iv*) additional actor attributes can be used and projected as supplementary variables. In the event subspace the interpretation is improved thanks to: *i*) item polarization; *ii*) special interpretation of event variance. All these features allow a very suitable detection of structural characteristics embedded in the affiliation matrix. Therefore the application of MCA is more appropriate and it is furthermore formally correct.

References

Borgatti, S., Everett, M.G., 1997. Network Analysis of 2-Mode Data. Social Networks **19**, 243–269

Greenacre, M., 1984. Theory and Applications of Correspondence Analysis. Academic Press

Roberts, J.M. Jr., 2000. Correspondence Analysis of Two-Mode Network Data. Social Networks **22**, 65–72

Wasserman, S., Faust, K., Galaskiewicz, J., 1989. Correspondence and Canonical Analysis for Relational Data. J. Math. Sociol. **1**(1), 11–64

Analysis of multivariate event networks

Jürgen Lerner

1 Introduction

Many dynamic social networks encode time-stamped *interaction events* rather than relational states (such as friendship or esteem) between actors. Examples include networks of communication events (such as email or telephone-call networks), collaboration (e. g., co-authoring scientific articles or jointly editing wiki pages), or political interaction-events which are frequently used in studies of international relations. Especially, communication or collaboration facilitated by the Internet naturally gives rise to event networks that are in many cases automatically logged and, thus, enjoy the benefit of simple availability.

In many applications of event network analysis, an event $e = (a,b,t,w)$ is characterized by four components

- the **source actor** a initiating e,
- the **target actor** (addressee) b of e,
- the **time** t when e happens, and
- the **type** w encoding what happens.

In the applications that we envision here, the event type w is given by a continuous (real) variable, where a negative value indicates hostile events and positive values indicate friendly interaction.

In this paper we propose a parameterized model for sequences of dyadic, typed events on a fixed set of actors. The model combines elements from the analysis of multivariate survival data, e. g., Hougaard (2000) and social network analysis, e. g., Robins et al. (2007). The time and type of an event on a given dyad (a,b) is dependent on events that happened earlier on the same or other dyads. For instance, if b initiated hostile events towards a the probability of a hostile event from a to b might increase. For illustration, we apply our model to event data encoding political

Jürgen Lerner
University of Konstanz, e-mail: lerner@inf.uni-konstanz.de

interaction. We test several hypotheses about the effect of externally given covariates (such as geographic proximity, form of government, or capability ratio) as well as postulated dependencies between events on different dyads (such as *"actors tend to collaborate with the enemies of their enemies"*).

2 Model

Given a particular event $e = (a,b,t,w)$ from a (time-ordered) sequence of events $E = (e_1,\ldots,e_N)$ we assume that e is (only) dependent on those events from E that happened before t. To obtain a tractable model those previous events are aggregated dyadwise to a multivariate and weighted network G_e. Each dyad is characterized by two values encoding the past cooperative respectively hostile interaction in such a way that events lying further back in the past have a decreasing influence. The probability density of e is assumed to be conditionally independent on all other events, given the network of past events G_e.

Similar to the proposal in Brandes et al. (2009), the probability density of an event $e = (a,b,t,w)$ is decomposed into two factors, the first modeling the frequency of events on the dyad (a,b) and the second modeling the conditional event type, given that an event happens on (a,b). This leads to a different (and, as we argue, valuable) model for typed events than the one proposed in Butts (2008). In particular, the second component (i. e., the conditional event type) admits very efficient parameter estimation and leads to robust explanations for the occurrence of conflict vs. cooperation.

References

Brandes U, Lerner J, Snijders TAB (2009) Networks evolving step by step: Statistical analysis of dyadic event data. In: Proc. 2009 Intl. Conf. Advances in Social Network Analysis and Mining (ASONAM 2009), pp 200–205

Butts CT (2008) A relational event framework for social action. Sociological Methodology 38(1):155–200

Hougaard P (2000) Analysis of Multivariate Survival Data. Springer-Verlag

Robins G, Snijders TAB, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p^*) models for social networks. Social Networks 29:192–215

Contributed Session 25

Rankings and Preferences

A Monte-Carlo study to evaluate value-added models for institutions' rankings

Bruno Arpino, Roberta Varriale

1 Introduction

In many fields of applied research and policy evaluation it is of interest to rank institutions with regard to their performance. This is a widely addressed issue in the educational literature, where it is often required to rank schools or universities with respect to their effectiveness. The seminal paper by Aitkin and Longford (1986) describes the advantage of using multilevel regression models compared to one-level models. The quality of the ranking obtained through multilevel models depends on the validity of the assumptions underlying the multilevel regression model, that are similar to those used in ordinary multiple regression analysis, such as homoscedasticity and normal distribution of the residuals. While some Monte Carlo simulation studies have been carried out in order to evaluate the robustness of multilevel models with respect to the parameter estimates and their standard errors in case of violations of these assumptions (see e.g., Maas and Hox (2004)), we focus on the effect of different model misspecifications on the ranking quality. Furthermore, we compare the performance of random and fixed effect models.

2 Main results

Our simulations show that the ranking obtained through multilevel models is reliable only for extreme institutions. Consistently with multilevel literature (see, e.g.,

Bruno Arpino,
Department of Decision Sciences and DONDENA 'Carlo F. Dondena' Reaserch Centre on Social Dynamiscs, Via Roentgen 1 - 20135 Milano (Italy), e-mail: bruno.arpino@unibocconi.it

Roberta Varriale,
Department of Statistics 'G.Parenti', University of Florence, Viale Morgagni, 59 - 50134 Firenze (Italy) - Department of Methodology and Statistics, Tilburg University, P.O. Box 90153 5000 LE Tilburg (The Netherlands), e-mail: roberta.varriale@ds.unifi.it

Goldstein and Healy (1995)), we find that it is easier to reliably rank the institutions with extreme performances but it is hard to precisely rank the institutions with average performances. However, we do not think this is a reason to abandon the approach because extremely "bad" and "good" performing institutions are usually the most interesting for researchers and policy makers. Second, the effect of non-normal errors at the second level can be detrimental also to rank extreme institutions. In particular, a highly asymmetric distribution (e.g., Chi-square) of second-level residuals implies a good ranking quality only of one tail of the distribution and a rather poor quality of the other tail. A bi-modal distribution, on the contrary, produces a low ranking quality for both tails. This highlights the importance of testing for normality of residuals' distribution. With respect to the data structure, large sample sizes help to increase the ranking quality, while the number and size of clusters, *per se*, play a less important role. We also find that a large ICC facilitates the ranking. Finally, discrepancies in the between and within effects of covariates is a crucial point for the quality of the ranking. In all experimental situations, the multilevel model with cluster means, that allows the between and within effects of the covariates to be different, performs much better than the others. Only when the between and within effects are equal for all the covariates, the three models perform very similarly. These results highlight the importance, also for clusters' ranking, to take into account that within-cluster and between-cluster relationships can be very different when dealing with multilevel data structures. We plan to extend the current simulation work in several other directions. First of all, the role of shrinkage in unbalanced data structures will be studied. Another important development of the work consists in the proposal of simple data transformations (Box-Cox type) and non-parametric methods for the estimation of random effects, as tools to improve the ranking quality in the presence of non-normal 2-level residuals.

References

Aitkin M., Longford N. (1986). Statistical modelling in school effectiveness studies (with discussion). *Journal of Royal Statistical Society A*, 149, 1-42.

Goldstein H., Healy M.J.R. (1995). The graphical presentation of a collection of means. *Journal of Royal Statistical Society A*, 158, 175-177.

Maas C.J.M, Hox, J.J.(2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2), 127-137.

Ordinal models to assess media reputation

Paola Cerchiello, Paolo Giudici

1 Introduction

Reputation can be defined as how an entity (private or public) is perceived by each of its stakeholder groups and reputation risk as the risk that an event will negatively influence stakeholder perceptions. Since reputation involves intangible assets (public opinion, perception, reliability, merit), it is not simple to define and consequently to measure and to monitor the correlated risk. Thus only few strategies have been put in place in practice. The first formal definition of reputational risk is due to Basel Committee on Banking Supervision that in 1997 stated that "Reputational risk arises from operational failures, failure to comply with relevant laws and regulations, or other sources. Reputational risk is particularly damaging for banks since the nature of their business requires maintaining the confidence of depositors, creditors and the general marketplace". In this context, media coverage plays a key role in determining a company reputation. This often occurs when a company reputation has been significantly damaged by unfair attacks from special interest groups or inaccurate reporting by the media. A detailed and structured analysis of what the media are saying is therefore especially important, because the media shape the perceptions and expectations of all the involved actors. Nowadays, natural language processing technologies enable companies and their business intelligence departments to scan a wide range of outlets, including newspapers, magazines, TV, radio, and blogs. The above mentioned collection of textual data, aimed at measuring the reputation and the reputational risk of an Institution, motivates the development of appropriate statistical models for the analysis of such data.

Paola Cerchiello,
Department of Statistics and Applied Economics 'L. Lenti' University of Pavia,
e-mail: paola.cerchiello@unipv.it

Paolo Giudici,
Department of Statistics and Applied Economics 'L. Lenti' University of Pavia,
e-mail: giudici@unipv.it

2 Main results

We propose two parallel approaches rooted in the context of, respectively, non parametric and parametric statistics. The former allows us to employ a scorecard approach based on flexible indexes with the final aim of creating a ranking. Besides non parametric models, we need a parametric model whose estimation allows not only to describe and rank reputation, but also to predict and, therefore prevent, reputational risks. In particular we need a parametric model suited for ordinal variables, as most reputational data is typically available in such format. With regards to non parametric approach, Giudici in 2007 proposed to employ the median as a location measure for each ordinal distribution, and the normalized Gini index as an indicator of the "consensus" on such location measure.

On the other hand the parametric approach is based on the CUB/CUBB models, that are mixtures of two or three random variables (shifted Binomial r.v. and uniform r.v.) able to model ordinal variables effectively and to estimate the latent components that guide article author. The above mentioned models are applied to a reputation variable produced by an opinion mining (OM) tool that executes a sentiment classification of newspaper articles regarding 40 Italian company. It determines the attitude (a judgment or an evaluation) of the writer with respect to a given topic. The OM result pursues the following structure: 1 (very bad news), 2 (bad news), 3 (neutral news),4 (good news) and 5 (very good news).

Results, obtained on the basis of real and simulated data, show not only an evident coherence between the two approaches, but also the power of the proposed methodology. In fact, on one hand we obtain a ranking of the companies based on the reputation emerging from newspaper articles, on the other hand we can evaluate and represent the latent components behind the textual data (for more details see Cerchiello and Giudici, 2010).

References

Giudici P. (2007). Governo dei Rischi: Il ruolo dei modelli statistici. *Isituto Lombardo (Rend. Lett.)Edizioni Universitarie di Lettere Economia Diritto.*, Vol.141, 361–376.

Cerchiello P., Giudici p. (2010). Ordinal statistical models to assess Reputational risk. *Technical Reports*, Quaderni di Dipartimento, n 67.

Inference on the CUB model: an MCMC approach

Laura Deldossi, Roberta Paroli

1 Introduction

The CUB model has been recently introduced in statistical literature by D'Elia and Piccolo (2005) and Iannario and Piccolo (2009) to model ordinal data expressing the preferences of raters within items or services. The model is defined as a finite mixture of two different distributions: a shifted Binomial and a discrete Uniform.

Bayesian analysis of the CUB model naturally comes from the elicitation of some priors on its parameters. In this case the parameters estimation cannot be performed via the classical EM method, but it must be obtained through the analysis of the posterior distribution. In the theory of finite mixture models complex posterior distributions are usually evaluated through computational methods of simulation (Früwirth-Schnatter, 2006), such as the Markov Chain Monte Carlo (MCMC) algorithms. Since the mixture type of the CUB model is non-standard, an MCMC algorithm has been developed here and its performance has been evaluated via simulation experiments.

2 Main results

Let r, with $r \in \{1,\ldots,m\}$, be the rank assigned by a rater to a given item of a preferences test. The discrete random variable R may be modeled as a mixture of a shifted Binomial(m,ξ) and a discrete Uniform(m), with $\xi \in [0,1]$ and $m > 3$, due to the identifiability conditions; the weight of the mixture is $\pi \in (0,1]$.

In the context of the preference analysis, the Uniform component may express the degree of uncertainty in judging an object on the categorical scale, while the

Laura Deldossi and Roberta Paroli,
Dipartimento di Scienze Statistiche, Università Cattolica SC, Milano.
e-mail: laura.deldossi@unicatt.it & roberta.paroli@unicatt.it
Partially supported by MIUR project PRIN2008 on "Modelli per variabili latenti basati su dati ordinali: metodi statistici ed evidenze empiriche" (Research Unit of Naples Federico II)

shifted Binomial component may represent the behavior of the rater with respect to a liking/disliking feeling for the object under evaluation. For any items we are interested in estimating the parameters ξ, that is a proxy of the rating measure, and π, that is inversely related to the uncertainty in the rating process.

If we introduce covariates of each subject i, for $i = 1,\ldots,n$, to better explain uncertainty and feeling parameters, π_i and ξ_i, the general formulation of a CUB(p,q) model is expressed by:

$$P(R_i = r; Y_i, W_i) = \pi_i \binom{m-1}{r-1} (1-\xi_i)^{r-1} \xi_i^{m-r} + (1-\pi_i)\frac{1}{m} \tag{1}$$

where $r = 1,2,\ldots,m$, Y_i and W_i are the covariates row vectors of dimension $p+1$ and $q+1$, respectively, linked to π_i and ξ_i by logit models:

$$\pi_i = \tfrac{1}{1+\exp^{-Y_i\beta}}; \;\; \xi_i = \tfrac{1}{1+\exp^{-W_i\gamma}}. \tag{2}$$

The parameters to be estimated are then the column vectors β and γ. In a Bayesian perspective we place independent priors on the parameters: we assume that each entry of vector β is Normal with known hyperparameters μ_B and σ_B^2 $\left(\beta_j \sim \mathcal{N}(\mu_B, \sigma_B^2\right)$, for any $j = 1,\ldots,p+1$; each entry of vector γ is Normal with known μ_G and σ_G^2 $\left(\gamma_j \sim \mathcal{N}(\mu_G, \sigma_G^2\right)$, for any $j = 1,\ldots,q+1$.

Bayesian inference will be executed by sampling from the posterior density through an MCMC algorithm, with the following two Metropolis-Hastings steps at the generic k-th iteration:

[1] the parameters $\beta_j^{(k)}$, for any $j = 1,\ldots,p+1$, are independently generated from a random walk and accepted or rejected according to the acceptance probability;

[2] the parameters $\gamma_j^{(k)}$, for any $j = 1,\ldots,q+1$, are independently generated from a random walk and accepted or rejected according to the acceptance probability.

At the end of a number N (large enough) of iterations we obtain N-dimensional samples of the parameters values that will be used to estimates each β_j and γ_j through the ergodic means.

Some simulation results will be used to check the statistical performance of the algorithm.

References

D'Elia A., Piccolo D. (2005). A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49, 917-934.

Früwirth-Schnatter S. (2006). *Finite Mixture and Markov Switching Models*, New York: Springer.

Iannario M., Piccolo D. (2009). A program in R for CUB models inference, Version 2.0. avaliable at *http://www.dipstat.unina.it*.

Contributed Session 26

Correspondence Analysis and Related Methods II

Comparing mental maps: Obama vs. McCain

Simona Balbi, Michelangelo Misuraca, Emma Zavarrone

1 Introduction

Network Text Analysis (NTA) has been developed for encoding the relationships among words in a text (Popping, 2000). This approach is based on the assumption that both language and knowledge can be modelled as networks by considering words as vertices and the relations among them as edges. NTA can be carried out with several methods.

Hereafter we will refer to Map Analysis, as in Diesner and Carley (2004). Map Analysis systematically extracts and analyses the links between words in a text in order to model the authors' "mental map" as networks of words. A map encoding scheme focuses analysts on investigating the "meaning" of a text, because it detects the relationships among words, among topics, and between them.

From a methodological viewpoint, NTA is rooted in a Social Network Analysis framework: *concepts* are equivalent to nodes, while the link between two concepts, which is referred to as a *statement*, corresponds to an edge (or an arc). The union of all statements per texts forms a map (Carley, 1997), which is graphically equivalent to a network. This correspondence between SNA and NTA allows analysing texts with SNA measures. On the other hand, as pointed out by Bourdieu (deNooy, 2003), analysing association structures by means of factorial techniques like Correspondence Analysis allows to enrich the comprehension of a phenomenon, because the relations taken into account are "objective". Furthermore, translating a network in a set of coordinates allows to compare different mental maps.

Simona Balbi,
University of Naples - Federico II, e-mail: simona.balbi@unina.it

Michelangelo Misuraca,
University of Calabria, e-mail: michelangelo.misuraca@unical.it

Emma Zavarrone,
IULM University - Milan, e-mail: emma.zavarrone@iulm.it

2 Theoretical framework and proposal

The problem of comparing two or more texts has been faced in several frames, particularly when the texts are tightly related such as translations in different languages, the open-ended questions and the corresponding answers in a survey, in plagiarism, and so on. In particular it is possible to consider *parallel corpora* when exact translations are available (e.g. EU multilingual documents collections), or *comparable corpora* when the texts are semantically linked (e.g. multinational surveis or political speeches in electoral campaign). In both cases it is necessary to align the texts in order to study structural similarities and differences.

The comparison of different texts by (generalised) Procrustes rotations has been proposed by Balbi and Esposito (2000) in a Textual Data Analysis frame, dealing with Italian advertisement campaigns of the same product in different periods, in order to explore language evolution. Similarly aiming at evaluating translations quality a *Cross-Language Correspondence Analysis* has been proposed by Balbi and Misuraca (2006). With a different standpoint Salem (2004) has discussed in a textometric frame the so called *textual resonance*, showing several interesting application fields.

This paper aims at proposing a study of Obama's and McCain's mental maps, analysing two *corpora* with a selection of the 2008 U.S. presidential election speeches for each "candidate". A separate *local correspondence analysis* (Aluja and Lebart, 1984) is performed on the adjacency matrix obtained from each *corpus*, than the latent semantic representations are compared by means of orthogonal Procrustes rotations. The final results will be reported elsewhere.

References

Aluja Banet T., Lebart L. (1984). Local and Partial Principal Component Analysis and Correspondence Analysis. In: COMPSTAT - Proceedings in Computational Statistics, Vienna: Physica Verlag, 113-118.

Balbi S., Misuraca M. (2006). Procrustes techniques for Text Mining. In: S. Zani, A. Cerioli, M. Riani, M. Vichi (eds). Data Analysis, Classification and the Forward Search, Heidelberg: Springer-Verlag, 227-234.

Balbi S., Esposito V. (2000). Rotated canonical analysis into a reference subspace. *Computational Statistics and Data Analysis*, 32, 395-410.

Carley K.M. (1997). Extracting Team Mental Models Through Textual Analysis. *Journal of Organizational Behavior*, 18, 533-538.

de Nooy W. (2003). Fields and networks: correspondence analysis and social network analysis in the framework of field theory. *Poetics*, 31, 305-327.

Diesner J., Carley K.M. (2004). *AutoMap1.2 - Extract, analyse, represent, and compare mental models from texts*. Technical Report. Carnegie-Mellon University.

Popping R. (2000). *Computer assisted text analysis*. London: Sage.

Assessing the response quality in ordered categorical data

Jörg Blasius

Responses to a set of items in survey data are not only associated with socio-demographic characteristics such as age, gender, and educational level, they are also associated with different kinds of response styles, such as acquiescence response style, extreme response style, and midpoint responding. Further, there are misunderstandings of questions, arbitrary responses, fatigue and other effects, which also reduce the quality of data. In general, when analyzing a battery of items, responses are related to the substantive concept, which we are mainly interested in, and to methodological effects. Applying categorical principal component analysis (CatPCA) to an item battery of survey data allows us to assess what part of the responses is due to substantive relationships and what part is attributable to methodological artifacts. In a first paper, Blasius and Thiessen (2009) demonstrated that the share of tied data in CatPCA can be used as a rough indicator for assessing the quality of data. This idea has been further developed so that we are now able to provide with a coefficient to describe the quality of responses in a given item set. Using different examples, we will show which part of variation can be explained by the substantive concept and which part is due to methodological induced variation.

References

Blasius J., Thiessen V. (2009). Facts and Artifacts in Cross-National Research: The Case of Political Efficacy and Trust. In: Haller M., Jowell R., Smith T.W. (eds.). *The International Social Survey Programme, 1985-2009. Charting the Globe*, London: Routledge, 147-169.

Jörg Blasius,
University of Bonn, Institute for Political Science and Sociology, Lennstr. 27, 53113 Bonn, Germany, e-mail: jblasius@uni-bonn.de

Adaptive factorial clustering for binary data

Alfonso Iodice D'Enza, Francesco Palumbo

1 Introduction

Tandem analysis (Arabie and Hubert, 1994) is a two-step technique aiming to find groups in high dimensional and sparse data: it combines factor analysis and clustering. *Factor K-means* proposed by Vichi and Kiers (2001) is an enhancement of tandem analysis where the two steps are iterated up to a stable solution. In the categorical data framework, Hwang *et al.* (2006) and Palumbo and Iodice D'Enza (2010) proposed a suitable combination of clustering with multiple correspondence analysis and non-symmetric correspondece analysis, respectively.

According to Palumbo and Iodice D'Enza's approach, this contribution presents a dynamic clustering procedure for high dimensional binary data that are arranged into subsequent batches: the first data batch is used to determine a 'starting' solution that is updated as further data batches are processed. The proposed method aims to cope with a two-fold problem: *i)* clustering very large data sets or data produced at a high rate (data flows), when it is convenient or necessary to process it in different 'pieces'; *ii)* if the data refer to different occasions or positions in space, when a comparative analysis of data stratified in chunks can be suitable.

2 Main results

The proposed strategy consists of a two-step iterative clustering procedure of binary row vectors: at each iteration a quantification of the starting binary variables is obtained so that the group structure underlying the statistical units is emphasized.

Alfonso Iodice D'Enza,
Università di Cassino e-mail: iodicede@gmail.com

Francesco Palumbo,
Università degli Studi di Napoli Federico II, e-mail: fpalumbo@unina.it

Consider a starting data batch consisting of n statistical units described by p binary attributes Z_j ($j = 1, \ldots, p$) and a set of K indicator variables X_k ($k = 1, \ldots, K$) assigning each statistical unit to one of the K groups. The aim of the procedure can be formalised as

$$(p_{X_1}, p_{X_2}, \ldots, p_{X_K}) : \sum_{j=1}^{p} E\left[P(X_k \mid Z_j) - P(X_k)\right] = \max! \tag{1}$$

The starting solution is obtained alternating iteratively two steps and it provides an optimal allocation of the units into the K groups and a low-dimensional quantification of the Z_j attributes.

When a new data batch comes in, the dynamic procedure runs as follow: the upcoming statistical units are first assigned to the K groups according to the former quantification of attributes; afterwards, the attribute quantification is updated according to the new allocation structure. It is worth to highlight that new data batches must have the same set of attributes Z_j and can consist of any amount of statistical units, even a single one.

Note that the updating process does not require to keep in memory the previous data structures. The information needed to update the solution refers to the attribute distributions: in particular, it takes to keep in memory the conditional attribute distributions within each of the K groups and the marginal distribution of the attributes. A similar dynamic update of data structures by keeping in memory the marginal distribution only has been proposed by Iodice D'Enza and Greenacre (2010).

References

Arabie P., Hubert L. (1994). Cluster analysis in marketing research. *IEEE Trans. on Automatic Control* AC. 19: 716–723.

Hwang H., Dillon W. R., Takane Y., (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. *Psychometrika* . 71, 161–171.

Iodice D'Enza A., Greenacre M. J. (2010). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In: proceedings of SIS09 *Statistical Methods for the analysis of large data-sets*. Pescara, Italy (submitted).

Palumbo F., Iodice D'Enza A. (2010). A two-step iterative procedure for clustering of binary sequences. In: *Data Analysis And Classification*. Springer, Heidelberg, 50–60.

Vichi M., Kiers H. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis* 37(1): 49–64.

Contributed Session 27

Methodology and Applications of Latent Class and Mixture Models

Statistic cognitive survey on Passito wine in Veneto region (Italy) from the consumer's point of view

Rosa Arboretti Giancristofaro, Stefano Bonnini, Elisa Grossule, Susanna Ragazzi, Luigi Salmaso

Iannario and Piccolo (2007) have recently proposed a mixture distribution, named CUB, for ordinal data. The use of such mixture distribution for modeling ratings is justified by the following consideration: the judgment that a subject expresses is the result of two components, uncertainty and selectiveness. The possibility of relating the parameters of CUB models to covariates makes the formulation interesting for practical applications.

This work presents such approach to study, evaluating and measuring the trend of the demand of Passito Wine in Veneto in order to define the target consumer profile. In this case study, a sample of 386 Passito wine consumers were interviewed. With this data-set, CUB model split consumers according to their preferences in different segments. Using CUB Models, the distribution of preferences expressed by the consumer is synthesized, through the Likert Scale and starting from the qualitative judgment, through a comparison among the entire distribution of the expressed valuations/judgments, and allows to place such comparison in an inferential context and finally to connect the valuation to the characteristics of the subjects.

Rosa Arboretti Giancristofaro,
Department TESAF, University of Padova, e-mail: rosa.arboretti@unipd.it

Stefano Bonnini,
Department of Mathematics, University of Ferrara e-mail: stefano.bonnini@unife.it

Elisa Grossule
Department TESAF, University of Padova, e-mail: elisa.grossule@unipd.it

Susanna Ragazzi
Department TESAF, University of Padova, e-mail: susanna.ragazzi@unipd.it

Luigi Salmaso
Department of Engineering and Management, University of Padova e-mail: salmaso@gest.unipd.it

References

D'Elia A., Piccolo D. (2005). A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49(3), 917-934.

Greene W. H., Hensher D. A., Rose J. M. (2007). Applied Choice Analysis: A Primer (2007). *Psychometrika*, 72(3), 1-711.

Iannario M., Piccolo D. (2007). A new Statistical Model for the Analysis of Customer Satisfaction (2007). *Quality Technology & Quantitative Management*, 7(2), 149-168.

Kuhfeld W. F. (2009). *Discrete Choice*, support.sas.com, 263- 618..

Covariate effects in multivariate latent growth models for the analysis of undergraduated student performances

Silvia Bianconcini, Silvia Cagnone and Paola Monari

The evaluation of the formative process in the University system has been assuming an ever increasing importance in the European countries. Within this context the study of student performance and capabilities plays a fundamental role. In order to analyze a cohort of students enrolled at the University of Bologna, Bianconcini and Cagnone (2009) have proposed a new general class of models, that combines: *i)* multivariate latent curves that describe the temporal behaviour of the responses, and *ii)* a factor model that specifies the relationship between manifest and latent variables.
This methodology was motivated by the data coming from the DWH of the University of Bologna. In particular, the authors focused on student achievements measured through two items observed over time. These latter were characterized by the interrelationship due to their temporal pattern as well as the interrelationship due to the presence of a latent construct. The proposed approach allows to explain the former by means of multivariate latent curves and the latter by means of a factor model simultaneously. Based on the Generalized Linear and Latent Variable Model (GLLVM) framework, the response variables are assumed to follow different distributions of the exponential family, with item-specific linear predictors depending on both latent variables, and measurement errors.
In this paper we extend this class of multivariate latent growth models to allow for direct covariate effects on the manifest variables and covariate effects on the latent variables. A full maximum likelihood estimation method is used to estimate all the model parameter simultaneously.

Silvia Bianconcini,
Department of Statistics, University of Bologna, e-mail: silvia.bianconcini@unibo.it

Silvia Cagnone,
Department of Statistics, University of Bologna, e-mail: silvia.cagnone@unibo.it

Paola Monari,
Department of Statistics, University of Bologna, e-mail: paola.monari@unibo.it

References

Bianconcini S. and Cagnone S. (2009), “A general multivariate latent growth model with applications in student careers Data warehouses”, under review.

On using item features to estimate parameters in IRT models

Mariagiulia Matteucci, Stefania Mignani, and Bernard P. Veldkamp

1 Introduction

In educational and psychological testing, the use of item response theory (IRT) has become widespread. On the basis of a mathematical model expressing the conditional response probability as a function of the latent trait underlying the response process, IRT allows the simultaneous estimation of item parameters and examinee ability.

Different estimation methods have been proposed in the literature, as marginal maximum likelihood (MML) or Bayesian estimation. However, all approaches are conditioned to the existence of a large number of respondents, which should ensure an accurate estimation of both item and person parameters. This is especially true in computerized adaptive testing (van der Linden and Glas, 2000), where large item banks are required in order to reduce item overexposure, to respect specific constraints in test assembly, and to provide a reliable estimate of the candidate's ability with a limited number of items. As a solution, the introduction of collateral information, when available, in the estimation process is possible. Matteucci, Mignani and Veldkamp (2009) have shown how background variables about individuals (e.g. scores on previous tests) could be introduced in the two-parameter normal ogive model estimation by using the Gibbs sampler (Geman and Geman, 1984) in a Bayesian framework.

In the present work, the introduction of prior information at the item parameter level is taken into account with the aim of reducing the efforts needed in the phase of test calibration.

Mariagiulia Matteucci, Stefania Mignani
Statistics Department "Paolo Fortunati", University of Bologna, Via Belle Arti 41, 40126 Bologna, ITALY, e-mail: m.matteucci@unibo.it, stefania.mignani@unibo.it

Bernard P. Veldkamp
Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS, e-mail: b.p.veldkamp@gw.utwente.nl

2 Main results

When the cognitive process underlying the responses is available, the prediction of item parameters like difficulty becomes possible (Irvine and Kyllonen, 2002). By using data from intelligence tests, where a set of explanatory variables concerning the item characteristics is available (e.g. type of arithmetic operation, numbers involved, etc.), the prediction of the real-valued item parameters in the two-parameter normal ogive model is implemented by using regression trees (Breiman, Friedman, Olshen, and Stone, 1984).

The results are twofold. Firstly, it is shown how predicted values from the regression tree can be used to set the parameters of informative prior distributions on the item parameters by adopting a Bayesian approach, and particularly using the Gibbs sampler in the Markov chain Monte Carlo (MCMC) methods. A better initial approximation of item parameters can especially be useful to reduce the number of candidates necessary to correctly calibrate the items. Secondly, predictions from the regression tree can be used directly in computer adaptive testing, as approximations of item parameters. While the common practice in adaptive testing is to consider known and fixed item parameters, uncertainty about item parameters is now introduced. An important consequence of using this approach is that a larger number of items is required to obtain accurate estimates of candidate ability. However, the substitution of estimated item parameters allows the omission of the item calibration process. Of course, the effective possibility of implementing this approach relies on the availability of high informative variables about the cognitive process needed to solve the items.

References

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and regression trees*, Belmont, CA: Wadsworth.

Geman S., Geman D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Irvine S.H., Kyllonen P.C. (2002). *Item generation for test development*, Mahwah, NJ: Lawrence Erlbaum Associates.

Matteucci M., Mignani S., Veldkamp B.P. (2009). Issues on item response theory modeling. In: Bini M., Monari P., Piccolo D., Salmaso L. (eds.), Statistical Methods for the Evaluation of Educational Services and Quality of Products, series "Contributions to Statistics", 29-45. Berlin-Heidelberg, Springer-Verlag.

van der Linden W.J., Glas C.A.W. (2000). *Computerized adaptive testing: theory and practice*, Boston, MA: Kluwer Academic Publishers.

Contributed Session 28

Markov and Graphical Modeling

Model selection in latent Markov models: a simulation study

Michele Costa, Luca De Angelis

1 Introduction

The major aim of this paper is the analysis of model selection procedures in latent (hidden) Markov models (LMM). In order to evaluate and compare existing proposals, we provide a Monte Carlo study which allows a powerful insight on the behaviour of the most widespread information criteria.

Although model selection has been deeply analyzed for both LMM (Cappé et al., 2005) and other mixture model types (Nylund et al., 2007), this topic is still an unresolved methodological issue. At the moment, there is not one commonly accepted statistical indicator for deciding on the number of latent states of the unobserved Markov chain. Furthermore, to our knowledge, there are no studies about the reliability and the precision of model selection instruments for LMM. Thus, the purpose of this paper is to fill this gap and provide a contribution to the debate on model selection in latent variable models.

The consistent identification of the number of latent states is a fundamental prerequisite to LMM parameter estimation. However, in many empirical applications of LMM modeling, no clue about this number is available. Therefore, using the LMM for exploratory purposes implies the selection of the cardinality of the discrete latent stochastic process underlying the observed time series. In particular, the number of latent states K of the Markov chain should be chosen in order to enable the model to account for the dynamic pattern of the observed time series.

Since the likelihood function is increasing for K, adding more latent states always improves the fit of the model. However, this improvement has to be traded off against the quadratic increase in the number p of parameters that have to be estimated by including a penalty term, usually specified as a function of p only or as a

Michele Costa,
Dipartimento di Scienze Statistiche, Universita' di Bologna, Italy, e-mail: michele.costa@unibo.it

Luca De Angelis,
Dipartimento di Scienze Statistiche, Universita' di Bologna, Italy, e-mail: l.deangelis@unibo.it

function of both p and the number of observations T. Thus, the order of a LMM is usually chosen considering one (or more) information criterion; in the following, we refer to the Bayesian information criterion (BIC), the Akaike information criterion (AIC), and two variants of the latter, the AIC3 and the Consistent AIC (CAIC).

2 Main results

As first step of our analysis, we define the probability functions of the LMM characterized by a homogeneous first-order Markov chain with K^* latent states and stationary distribution. In order to simulate the data set for one categorical variable X_t, with $t = 1, ..., T$, we set the conditional probabilities $r_{ij} = P(X_t = i|S = j)$, which indicate the probability of observing a particular value i of variable X_t, for $i = 1, ..., n$, given the membership of the observation to the latent state j, for $j = 1, ..., K^*$. We propose the values $n = 3$, 4, 6, and $T = 100$, 500, 1000 and simulate different data sets for each combination.

As second step of our analysis, we estimate LMMs with different number of latent states K for each simulated data set and we compute four information criteria: AIC, BIC, AIC3, and CAIC. Hence, we compare the value of K suggested by these criteria to the true K^*.

We are able to analyze the behaviour of information criteria with respect to some critical characteristics which, in our study, are represented by n, T, and the state-dependent conditional probabilities r_{ij}.

Simulation results show that the four information criteria tend to underestimate the number of latent states for lower values of T. Furthermore, underestimation of K seems, also for larger values of T, less relevant for AIC and AIC3 than for BIC and CAIC.

A crucial issue is represented by the state-dependent conditional probabilities r_{ij}. Increasing the level of uncertainty about the membership of the value i to a particular latent state j significantly reduces the power of the information criteria. Finally, all information criteria seem to be quite robust with respect to different values of n and, consequently, to the number of parameters p.

References

Cappé O., Moulines E., Rydén T. (2005). *Inference in Hidden Markov Models*, New York: Springer.

Nylund K.L., Asparouhov T., Muthén B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14 (4), 535-569.

Bayesian analysis of longitudinal categorical data via latent Markov models

Silvia Pandolfi, Francesco Bartolucci and Alessio Farcomeni

1 Introduction

The latent Markov (LM) model was introduced by Wiggins (1973) for the analysis of longitudinal categorical data, when the interest is in describing the evolution of a latent characteristic of a group of individuals over time. The basic assumption of this model is that the response variables are conditionally independent given a latent process which follows a first-order Markov chain. The initial version of the LM model of Wiggins (1973) was extended in several ways on the basis of more interpretable parametrizations and in order to include individual covariates. For a general overview see Bartolucci et al. (2010). The literature on LM models is strongly related to that on hidden Markov (HM) models for time series.

Though Bayesian inference for HM models is widely developed (see, among others, Robert et al., 2000), the same cannot be said for LM models. However, when applied to these models, the Bayesian approach has some advantages with respect to the Maximum Likelihood approach. The main advantage is that it allows for formal assessment of the number of regimes (or states), which is an important issue in every application, and makes model averaging straightforward.

In this work, we propose a Bayesian approach for LM models which is based on the reversible jump (RJ) algorithm (Green, 1995) and its continuous time version (Stephens, 2000). The approach is illustrated through a series of applications involving socio-economic data. More details on this approach are given in the following section.

Silvia Pandolfi, Francesco Bartolucci
Department of Economics, Finance, and Statistics, University of Perugia,
e-mail: pandolfi@stat.unipg.it, bart@stat.unipg.it

Alessio Farcomeni
Department of Experimental Medicine, Sapienza - University of Rome,
e-mail: alessio.farcomeni@uniroma1.it

2 Main results

First of all we propose a system of priors for the parameters of the different versions of the LM model, which may or may not include individual covariates. If available, these covariates may affect the distribution of the latent process or the conditional distribution of the response variables given this process. When the parameters of interest are probabilities, e.g. initial or transition probabilities of the latent process, we use a system of priors based on Gamma distributions. When these parameters are expressed on a different scale, e.g. regression coefficients involved in a logit parametrization of the response probabilities, we rely on Normal distributions.

Once the prior distribution of the model parameters has been chosen, we estimate their posterior distribution, together with the posterior distribution of the number of states, by an RJ-MCMC algorithm (Green, 1995), which is based on alternating two different moves. The first move, based on a Gibbs or a Metropolis-Hastings (MH) sampler, is aimed at drawing the model parameters conditionally on the number of states. The second move is transdimensional and also draws the number of states.

For the more complex models, the above strategy is compared with a continuous time version of the RJ algorithm due to Stephens (2000). Once the parameters of the model have been drawn through a Gibbs or MH sampler, the approach consists of performing a birth or a death move with a certain probability. The chosen move is always accepted and the time to the next move is simulated from a suitable distribution. The posterior probability of a certain number of states is computed as a quantity proportional to the overall permanence time in the corresponding model. In the RJ algorithm, instead, this estimated probability is proportional to the number of visits of this model.

The comparison between the two algorithms is performed under different situations in terms of model complexity, sample size, and presence/absence of covariates.

References

Bartolucci F., Farcomeni A., Pennoni F. (2010). An overview of latent Markov models for longitudinal categorical data. *Technical Report, arXiv:1003.2804*.

Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

Robert C., Ryden T., Titterington D. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, series B*, 62, 57–75.

Stephens M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics*, 28, 40–74.

Wiggins L.M. (1973). *Panel analysis. Latent probability models for attitude and behavior processes.* Elsevier Scientific Pub. Co., Amsterdam, New York.

On the identification of discrete graphical models with hidden nodes

Elena Stanghellini, Barbara Vantaggi

1 Introduction

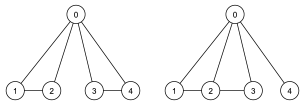
We focus on the identification of discrete undirected graphical models, (see Lauritzen, 1996, Ch.4), with one unobserved binary variable and establish a necessary and sufficient condition for the rank of the transformation from the natural parameters to the parameters of the model to be full. This ensures local identification of this class of models. These models generalize the latent class model, by allowing associations between the observed variables conditionally on the latent one. The practical importance of this issue is witnessed by several applied papers.

For non-full rank models, the obtained characterization allows us to find the expression of the (sub)space where the identifiability breaks down. Geometrically, this corresponds to the singularities in the parameter space (Drton, 2009). This in turn allows us (a) to derive a reparametrization that leads to an identified model and (b) to compute the correct dimension of the model. The condition is based on the topology of the undirected graph associated with the model and relies on the faithfulness assumption. The non-identifiability issue mentioned above has considerable repercussions on the asymptotic properties of standard model selection criteria (e.g. LRT, BIC), whose applicability and correctness may no longer hold (Drton, 2009). For non-full rank models, the asymptotic distribution of the LRT is also studied.

Elena Stanghellini,
D.E.F.S., Università di Perugia,
Via Pascoli, 1, I-06100 Perugia, e-mail: elena.stanghellini@stat.unipg.it

Barbara Vantaggi,
Dept. Metodi e Modelli Matematici, Università “La Sapienza”,
Via Scarpa 16, I-00161 Roma, e-mail: vantaggi@dmmm.uniroma1.it

2 The model



(a) (b)

Fig. 1 Two undirected graphs representing two possible models

Let $G^K = (K,E)$ be an undirected graph with $K = \{0,1,\ldots,n\}$, with 0 a node related to a binary unobserved variable A_0. Denote with G^0 the induced undirected graph of the observed variables. Let $l = \prod_v l_v$, with l_v the level of A_v, $v \in K \setminus 0$. Consider X to be the $2l \times 1$ vector of entries of the multidimensional contingency table and assume that the elements of X are independent Poisson variables with $E(X) = \mu_X$. Let $\log \mu_X = Z\beta$, where Z is a $2l \times p$ suitable design matrix; β is a p-dimensional vector of unknown parameters. Let $Y = LX$, with $L = (1,1) \otimes I_l$, the $l \times 1$ vector of the counts in the marginal table, marginal w.r. t. A_0. The elements of Y are independent Poisson random variables with $\mu_Y = Le^{Z\beta}$. By the inverse function theorem, a model is locally identified if the rank of the transformation from the natural parameters μ_Y to the new parameters β is full. In this context, this is equivalent to the rank of following derivative matrix

$$D(\beta)^T = \partial \mu_Y^T / \partial \beta = \partial (Le^{Z\beta})^T / \partial \beta = (LRZ)^T$$

being full, where $R = \mathrm{diag}(\mu_X)$. One of the main result is a necessary and sufficient condition for $D(\beta)$ to be full rank everywhere in the parameter space, that is:
(i) the undirected graph G^O of the observed variables O contains at least one m-clique C, with $m \geq 3$;
(ii) for each clique C_0 in G^O there exists a sequence in G^O $\{S_s\}_{s=1}^q$ of complete subgraphs such that (a) $|S_s| \leq |S_{s-1}|$ for $s \in \{1,\ldots,q-1\}$, $S_0 = C_0$ and $|S_q| = 1$; (b) for $s \in \{1,\ldots,q-1\}$ and for all $i \in S_s$ there exists $j \in S_{s+1}$ such that $(i,j) \notin E$. It follows that the model associated with the graph in Figure 1(a) is full rank, while the one associated with the graph in Figure 1(b) is not full rank. The asymptotic distribution of the LRT of this second model against the saturated one is also derived.

References

Drton M. (2009). Likelihood ratio tests and singularities. *Annals of Statistics* **27**, 2, 979-1012.

Lauritzen, S.L. (1996). *Graphical Models*, Oxford University Press: Oxford.

Contributed Session 29

Variable Structures

Hierarchical Factorial Classification of variables: methods and applications

Sergio Camiz, Jean-Jacques Denimal

1 Introduction

In this paper we propose some improvements of the *Hierarchical Factorial Classification* of variables (in the following *HFC*), based on the method proposed by Denimal (2000), and further discussed e.g. by Camiz et al. (2006).

The original method was first proposed to continuous characters and then extended to frequency data. In the continuous case, once standardized all original characters, each of them is considered a singleton group and the representative variable of this group. Then the iterated procedure for the construction of the hierarchy is the following:

1. a non-standardized *PCA* is performed on all pairs of representative variables, choosing the pair with least second eigenvalue
2. the two groups represented by the chosen pair are merged in a new group, that is a node of the hierarchy
3. the first principal component is taken as representative variable of the new group and the second as node variable, showing the internal differences of the node.
4. the second eigenvalue is taken as the node's hierarchy index
5. based on this 1mphPCA, both variables and units may be represented at each node on a factor plane

The procedure with frequency data is similar, but based on the correspondence analysis chi-square metrics.

Sergio Camiz,
Sapienza Università di Roma, e-mail: sergio.camiz@uniroma1.it

Jean-Jacques Denimal,
Université des Sciences et Technologies de Lille e-mail: jean-jacques.denimal@univ-lille1.fr

2 Main results

Recently, the method has been improved in several directions Denimal (2007a), Denimal (2007b).

1. The initial hierarchy obtained by the previous method is improved through an optimization process, that may be interpreted as a k-means type procedure among hierarchies in order to maximize the internal coherence of the nodes, in particular at the upper levels.
2. In order to identify a suitable cutpoint, for each variable a forward multiple regression is performed, in order to check to what extent it may be described through the set of the node variable of the class to which it belongs and of all nodes upper to that class. Thus, a suitable partition is the one in which all the upper nodes contribute significantly to the regression of at least one variable.
3. A segmentation of the units, based on this pruned and optimized hierarchy is proposed in optimized form. Indeed, starting from the top of the hierarchy, at each step, a partition of the units in two sub-clusters is obtained through a k-means technique and a mean test is performed to check for the splitting significance.
4. The methodology has been adapted to binary tables crossing units and characters levels within the frame of multiple correspondence analysis Denimal (2008). This is performed through a previous optimized quantification of each level, to which further the described technique is applied. The quantification stage of a level c of a character is obtained from the correspondence analysis of the table crossing the two levels $\{c, \overline{c}\}$, with $\overline{c}$ the complement of c, with all the level of the other characters. Then as scores of the units, the coordinates of their projections on the factor extracted by that correspondence analysis are taken.

   The new procedures will be introduced through examples.

References

Camiz S., Denimal J.J., Pillar,V.D. (2006). Hierarchical factor classification of variables in ecology. *Community Ecology* 7 (2), 165-179.

Denimal J.J. (2000). Correspondances hiérarchiques: une nouvelle approche. In: XXXIIiémes Journées de Statistiques, 15-19 mai 2000. Fès (Maroc).

Denimal J.J. (2007). Classification factorielle hiérarchique optimisée d'un tableau de mesures. *Journal de la Société Françcaise de Statistique*, 148 (2), 29-63.

Denimal J.J. (2007). Classification factorielle hiérarchique optimisée des lignes et des colonnes d'un tableau de contingence. *Journal de la Société Française de Statistique*, 148 (3), 37-70.

Denimal J.J. (in press). Extension aux correspondances multiples de la classification hiérarchique optimisée. Accepted in *Journal de la Société Française de Statistique*.

Sensitivity analysis of composite indicators through Mixed Model Anova

Cristina Davino, Rosaria Romano

1 Introduction

The aim of the paper is to develop a new approach for the analysis of Composite Indicators (CI) in the theoretical framework of explorative and confirmative analysis.

It is a matter of fact that the requirement to synthesize univariate indicators by means of a CI is becoming more and more common in all those contexts where the interesting phenomenon cannot be directly observed and measured due to the presence of several and different concurrent factors. Once a CI is constructed, a post-analysis of its stability is advisable before employing it in a decision process. The values of a CI and/or the ranking deriving from a CI depend on the methodological choices faced in its construction. These choices are well known in literature as uncertainty factors (Nardo et al., 2005) and they involve all the steps followed in the CI definition process: definition of the phenomenon to be measured (selection of factors, indicators and statistical units), pre-processing of the original indicators (missing data imputation, indicators transformations), construction of the CI (identification of the system of weights, selection of the aggregation method).

The paper is embedded in the Sensitivity Analysis (SA) (Saltelli et al., 2008) framework where the aim is to identify the contribution of each uncertainty factor on the obtained CI. In literature SA is mainly investigated computing a variance based sensitivity measure for each uncertainty factor. The proposal of the present contribution is to present an alternative CI Sensitivity Analysis based on a combination of Mixed Model Analysis of Variance models (McCulloch and Searle, 2001) and multivariate methods. Besides the evaluation of the impact of the uncertainty factors on the construction of the CI, the proposed approach allows to highlight the individual differences among the observations as well.

Cristina Davino,
University of Macerata, e-mail: cdavino@unimc.it

Rosaria Romano,
University of Macerata, e-mail: rosaria.romano@unimc.it

2 The proposed approach

Analysis of Variance (ANOVA) is a very useful method in cases where the objective is an assessment of the impact of some controllable factors (categorical variable) on a specific response (continuous variable). The impact is significant if the variability *between* the groups defined by the factor levels (categories) is much larger than the variability *within* the groups. Let X (N $\times$ P) be a data matrix of P indicators observed on N observations (for example countries) and let's consider for simplicity only three uncertainty factors: indicators transformation (t), weighting (w) and aggregation method (a), respectively with I, J and K levels. The model can be written as:

$$y_{ijkn} = \mu + t_i + w_j + a_k + c_n + tw_{ij} + ta_{ik} + wa_{jk} + ct_{ni} + cw_{nj} + ca_{nk} + e_{ijkn} \quad (1)$$

where y_{ijkn} is the n^{th} observation obtained using the i^{th} (i=1,...I) level of t, the j^{th} (j=1,...J) weighting scheme and the k^{th} (k=1,...K) aggregation method. In the previous model, the general mean is represented by μ, while t_i, w_j, a_k are the main effects of the three uncertainty factors and tw_{ij}, ta_{ik}, wa_{jk} are their interaction effects. These are all fixed factors. The main effect of the extra factor represented by the countries is c_n, while ct_{ni}, cw_{nj} and ca_{nk} are the interactions between countries and uncertainty factors and e_{ijkn} is the random error. As these countries can be viewed as one specific 'sample' of the whole population of countries, the related factor is a random factor. An ANOVA model including both fixed and random factors is called Mixed Model ANOVA. Results from model (1) show which uncertainty factors strongly affect or not the stability of the CI and also the impact of these effects on each single country. In order to better explore such differences and similarities among the countries, a Principal Component Analysis exploiting all the advantages of the factorial methods is performed on the residuals of an ANOVA model without the countries factor.

References

McCulloch C.E., and Searle S.R. EDS (2001). *Generalized, linear, and mixed models*, United States of America: John Wiley & Sons, Inc.

Nardo M., Saisana M., Saltelli A., Tarantola S., Hoffman A., Giovannini E. EDS (2005). *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Statistics Working Papers 2005/3, OECD, Statistics Directorate.

Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. EDS (2008). *Global Sensitivity Analysis. The Primer* , England: John Wiley & Sons Ltd.

A model for the clustering of variables taking into account external data

Karin Sahmer

1 Introduction

A method for the clustering of variables (CLV) was proposed by Vigneau and Qannari (2003). This method is based on a hierarchical clustering followed by a partitioning algorithm and includes several options. It can be used when variables with a negative correlation should be grouped together. It is also possible to use this method when a negative correlation between variables shows disagreement. In both cases, it is possible to take into account external data in the clustering procedure.

Sahmer (2006) analysed the CLV method in the case where a high negative correlation shows a proximity of variables. In that paper the clustering without taking into account external data was considered. The present communication concerns the case of grouping together only variables with positive correlations, taking into account external data. An important application of this option is the segmentation of consumers according to their liking of products, taking into account sensory data.

A model for the clustering of variables taking into account external data is proposed. The corresponding covariance matrix is specified. Then, the properties of the CLV method used on this covariance matrix are analysed.

2 Main results

The proposed model concerns the existence of K groups of variables $G^{(1)}, \ldots, G^{(K)}$. The number of variables in group $G^{(k)}$ will be denoted by $p^{(k)}$. In each group, the variables depend on some external variables. This relationship is assumed to be linear. So a variable $x^{(k)}$ belonging to group $G^{(k)}$ can be expressed as a linear combination $\mathbf{z}'\beta^{(k)}$ of the external variables $z_1, z_2, \ldots, z_q$ plus en error term ε. The random

Karin Sahmer,
Groupe ISA, e-mail: k.sahmer@isa-lille.fr

vector corresponding to the q z-variables is denoted by $\mathbf{z}$. So the equation for the j^{th} variable belonging to group $G^{(k)}$ is given as follows:

$$x_j^{(k)} = \mathbf{z}'\beta^{(k)} + \varepsilon_j^{(k)} \tag{1}$$

It is assumed that all error terms are uncorrelated with each other. The variance of the error terms is supposed to be equal: $\text{var}\left(\varepsilon_j^{(k)}\right) = \sigma^2 \; \forall j,k$. Furthermore, there is no correlation between the z-variables and the error terms. The covariance matrix of $\mathbf{z}$ will be denoted by $\Sigma_\mathbf{z}$.

Under this model, the covariance matrix of $\mathbf{x}^{(k)}$ (the random vector of the variables belonging to group $G^{(k)}$) is equal to:

$$\Sigma_\mathbf{x}^{(k)} = \mathbf{1}_{p^{(k)}}\mathbf{1}'_{p^{(k)}}\beta^{(k)'}\Sigma_\mathbf{z}\beta^{(k)} + \sigma^2\mathbf{I} \tag{2}$$

where $\mathbf{1}_{p^{(k)}}$ is the vector consisting of $p^{(k)}$ ones and $\mathbf{I}$ is the $p^{(k)}$-dimensional identity matrix.

The matrix of the covariances between variables of different groups $G^{(k)}$ et $G^{(m)}$ is given as follows:

$$\Sigma_\mathbf{x}^{(k,m)} = \mathbf{1}_{p^{(k)}}\mathbf{1}'_{p^{(m)}}\beta^{(k)'}\Sigma_\mathbf{z}\beta^{(m)} \tag{3}$$

while the matrix of covariances between $\mathbf{x}^{(k)}$ and $\mathbf{z}$ is equal to:

$$\Sigma_\mathbf{xz}^{(k)} = \mathbf{1}_{p^{(k)}}\beta^{(k)'}\Sigma_\mathbf{z}. \tag{4}$$

We will consider the clustering on the covariance matrix assumed above. When the CLV method taking into account external data is used, the decrease in the clustering criterion is equal to 0 when two subgroups of the same group $G^{(k)}$ are merged. When variables of different groups are merged, this decrease is larger than 0. Hence, the hierarchical algorithm of CLV, taking into account external data, will find the correct partition. It can further be shown that the hierarchical clustering without taking into account the external data also finds the correct partition. A simulation study will be performed in order to investigate the properties of the clustering algorithms performed on a sample covariance matrix.

References

Sahmer K. (2006). *Propriétés et extensions de la classification de variables autour de composantes latentes. Application en évaluation sensorielle.* PhD Thesis, Rennes, France, and Dortmund, Germany.

Vigneau E., Qannari E.M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics – Simulation and Computation*, 32 (4), 1131-1150.

Contributed Session 30

Risk Analysis

The distribution of the stochastic dominance index for risk measurement

Silvia Facchinetti, Paolo Giudici, Silvia Angela Osmetti

1 Introduction

The most employed approaches about operational risk (see Cruz, 2002), considers a quantitative approach and calculate the value at risk to derive the total economic capital required to protect an institution against possible losses. Figini and Giudici (2010) show that operational risk measurement is possible also for data in ordinal scale, and suggest as measure of risk the stochastic dominance index (SDI). Operational data for risk measurement are typically summarized in a matrix of J business lines and I event types. For each event type, in a specific business line, we have two different measures: the frequency and the severity expressed in an ordinal scale. To summarize them in a tendency measure, we structure a data set which counts, for each event type-business line and for a given severity, the absolute frequency. For the i-th event type and for the j-th business line, the SDI is:

$$SDI = \sum_{h=1}^{K} F_h / K \tag{1}$$

where K is the categories number of the severity categorical r.v. X and F_h $(h = 1, ..., n)$ are the cumulative frequencies of X.

In this paper we derive the distribution of SDI, thus allowing exact inference to be performed. Confidence intervals and testing rules are particularly useful in the context of operational risk, as they can help to prioritize and prevent operational failures in a quality control framework, so to effectively reduce the impact of risks

Silvia Facchinetti, Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, e-mail: silvia.facchinetti@unicatt.it

Paolo Giudici, Dipartimento di Statistica e Economia applicate, Università degli Studi di Pavia, e-mail: giudici@unipv.it

Silvia Angela Osmetti, Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, e-mail: silvia.osmetti@unicatt.it

ex ante and not ex post. We also derive the distribution of summary means of such measures for all business lines, particularly important in some applications.

2 Main results

Let X be an ordinal categorical r.v. with modality x_h for $h = 1, ..., K$, such that $x_h \preceq x_{h+1}\ \forall h$. Let also $(y_1, ..., y_j, \ldots, y_n)$ be an ordered random sample of size n from X with values $y_j = x_1, ..., x_K$ and let r_h $(h = 1, ..., K)$ be the number of the observations y_j in the sample equal to different values x_h. We now consider $r_1, ..., r_K$ as r.vs., where every r_h follows a binomial distribution with parameters $(n - \sum_{i=1}^{h-1} r_i; p_h/c_h)$, where p_h is the probability related to the modality x_h and $c_h = 1 - \sum_{i=1}^{h-1} p_i$. Let now

$$\begin{cases} u_i = r_i \qquad i = 1, ..., K-2 \\ u_{K-1} = (K-1)r_1 + (K-2)r_2 + \ldots + r_{K-1} + n = r_{K-1} + A_K. \end{cases} \tag{2}$$

The distribution of $K \cdot SDI$ is $p(u_{K-1}) = \sum_{u_1,...,u_{K-2}} p(\underline{u})$ with $\underline{u} = (u_1, ..., u_{K-1})$ and

$$p\ (\underline{u}) = \sum \left[\prod_{h=1}^{K-2} \binom{n - \sum_{i=1}^{h-1} u_i}{u_h} \left(\frac{p_h}{1 - \sum_{i=1}^{h-1} p_i} \right)^{u_h} \left(1 - \frac{p_h}{1 - \sum_{i=1}^{h-1} p_i} \right)^{n - \sum_{i=1}^{h} u_i} \right] \cdot$$

$$\cdot \binom{n - \sum_{i=1}^{K-2} u_i}{u_{K-1} - A_K} \left(\frac{p_{K-1}}{1 - \sum_{i=1}^{K-2} p_i} \right)^{u_{K-1} - A_K} \left(1 - \frac{p_{K-1}}{1 - \sum_{i=1}^{K-2} p_i} \right)^{n - \sum_{i=1}^{K-1} u_i} . \tag{3}$$

The sum is extended to all r_h such that $\underline{u} = (g_1(r_1, ..., r_{K-1}), \ldots, g_{K-1}(r_1, ..., r_{K-1}))$. We note that the distribution of SDI depends on the distribution law of the r.v. X. We analyze the SDI distribution assuming equi-distribution and degenerate distribution for X. Obviously, it is necessary to check the hypothesized distribution by a goodness of fit test. In particular we propose to modify the Kolmogorov-Smirnov test developed for discrete r.v. in Facchinetti and Osmetti (2009). Moreover, we assume that p_h are stochastic. Finally we conclude with an application to real data.

References

Cruz M. (2002). *Modeling, measuring and hedging operational risk*, New York: Wiley

Facchinetti S., Osmetti S.A. (2009). The Kolmogorov-Smirnov goodness of fit test for discrete extreme value distributions. In: Classification and Data Analysis 2009, Book of Short Papers, 485-488.

Figini S., Giudici P. (2010). Risk measures for ordinal variables, submitted. Discrete distributions, Boston: Houghton Mifflin Company.

Concentration measures for risk analysis

Silvia Figini, Paolo Giudici and Pierpaolo Uberti

1 Introduction

In this contribution we present a novel approach to employ concentration measures for financial risk management. We test how our proposal works on a real data set provided by a banking institution. Furthermore, we compare the results achieved following our approach in terms of operational risk capital requirement with respect to classical measures as the VaR (Value at Risk).

2 Main results

In this contribution we show the empirical evidences on operational risk capital requirement collected on a real data set. The data set reports operational risk data for the following business lines (see Basel, 2001): Commercial Banking, Retail Banking, Retail Brokerage and Trading and Sales. As event types the data set shows: Internal Fraud, External Fraud, Business disruption and system failures, Execution, delivery process management and Clients products and business practices.

Following the guideline of Basel 2, the event types may be categorised in terms of frequency (the number of loss events during a certain time period) and severity (the impact of the event in terms of financial loss).

In operational risk management, frequency and severity are regarded as random variables. For each business line and event type, the total loss is based on the convo-

Silvia Figini,
University of Pavia, e-mail: silvia.figini@unipv.it

Paolo Giudici,
University of Pavia e-mail: giudici@unipv.it

Pierpaolo Uberti,
University of Pavia e-mail: pierpaolo.uberti@unipv.it

lution between the frequency and the severity probability functions derived through Monte Carlo algorithm.

A crucial role in operation risk management measurement, is represented by the selection of an appropriate functional form for loss frequency and severity distributions.

Considering the frequency distribution, common choices in terms of probability distribution functions are binomial distribution, poisson and negative binomial distribution. On the other hand, in order to model the severity, on the basis of the quatitative nature of the data at hand, the functions selected in the applications are the log-normal and the gamma density functions.

The main objective in operational risk management is to cover unespected annual loss computing the operational risk capital requirement (ORR) given by the 99.9 percentile of annual loss (VaR) minus the mean annual loss (see e.g. Alexander, 2003).

In this contribution we will compare the results achived following Alexander (2003); Cruz (2002) and our approach based on concentration indexes.

In order to motivate our methodological proposal, we point out that the capital requirement can be calculated proportional to a concentration measure that summarises, for a given business line and event type, the distribution of the losses.

We employ the VaR obtained via Monte Carlo, as a measure of capital requirement and the Gini index as a concentration measure. On the basis of our proposal, the final estimation of the annual capital requirement is a combination between a measure of risk (VaR) and a concentration measure (Gini index).

References

Alexander C. (2003). Operational Risk: regulation, analysis and management, Prentice Hall.

Basel Committee on Banking Supervision (2001). Operational Risk, Consultative Paper.

Cruz M. (2002). Modeling, Measuring and Hedging Operational Risk, Wiley.

Gini C. (2001). Measurement of Inequality and Incomes. *The Economic Journal*, vol. 31, pp. 124-126.

Robust estimation and prediction for credit risk models

Silvia Figini, Luigi Grossi

1 Introduction

The main objective of this contribution is to compare classical predictive models and robust predictive models able to predict a binary target Y as a function of p covariates. To reach this objective, we have implemented a robust logistic regression model using forward search (see e.g. Atkinson and Riani, 2000) and a classical logistic regression. In order to choose the best model, we consider both a threshold independent criteria as well as a novel financial loss function (see e.g. Hand and Krzanowski, 2009). Empirical evidences are given on a real financial data provided by a German rating agency. In terms of out-of-sample performances, we find that robust models perform much better that classical models.

2 Methodological proposal

The empirical analysis is based on annual 1996–2004 data from Creditreform, which is one of the major rating agencies for SMEs in Germany, for about 1000 firms belonging to J different business sectors. While for a classical rating system only quantitative data are needed, our rating system want to focus on all types of information and data. Particularly, we use two types of data: quantitative data, in the form of accounting data provided by the companies themselves and qualitative data in the form of questionnaire, provided by the company themselves and later cleaned by financial analysts.

Silvia Figini,
University of Pavia, e-mail: silvia.figini@unipv.it

Luigi Grossi,
University of Verona, e-mail: luigi.grossi@univr.it

Our data set consists of a binary response variable Y_{itj} and a set of explanatory variables given by financial ratios, time dummies and analysts recommendations. Based on our dataset we present the following specification for a general class of predictive models for default estimation: for observation i, $(i = 1,...,n)$, time t, $(t = 1,...,T)$ and business sector j, $j = 1,...,J$, let Y_{itj} denote the response solvency variable and let X_{itj} denote a $p \times 1$ vector of candidate predictors. The elements of $Y_{itj} = (y_{1tj},...,y_{ntj})'$ are modelled as conditionally independent random variables from a simple exponential family:

$$\pi(Y_{itj}|X_{itj}) \propto exp\left\{\frac{Y_{itj}\theta_{itj} - b(\theta_{itj})}{a_{itj}(\phi)} + c(Y_{itj},\phi)\right\}, \tag{1}$$

where θ_{itj} is the canonical parameter related to the linear predictor $\eta_{itj} = X'_{itj}\beta$ with a $p \times 1$ vector of regression coefficient β, ϕ is a scalar dispersion parameter and a_{itj}, b, c are known functions with $a_{itj}(\phi) = \frac{\phi}{\omega_{itj}}$, where ω_{itj} is a known weight. We are interested in predicting the expectation of the response as a function of the covariates. In the case of a simple binary logit model, the expectation is just the probability that the response is 1:$E(Y_{itj}|X_{itj}) = \pi(Y_{itj} = 1|X_{itj})$ (see e.g. Dobson, 2002).

As classical models can be dramatically influenced by few outliers, we suggest to adopt a robust version of logit models. This goal can be achieved by getting unit weights which are inversely related to the degree of outlyingness of each observation. In this paper weights are computed comparing forward search trajectories (see Atkinson and Riani, 2000, for details) of deviance residuals with a given threshold based on their approximate distribution along the search. In this way, the masking effect, which is the main drawback of classical outlier detection methods, is avoided. Usually outliers which could be, for instance, good rated firms with a bad financial situation, are down-weighted and the predictive accuracy of the classification rule is improved.

Summarizing our contribution, we have compared classical logistic regression and robust logistic regression in order to estimate the probability of default of a set of SMEs. Empirical evidences collected on this data show that robust logistic regression is the best model to estimate the probability of default. In our opinion the results achieved should be improved using robust time-dependent model as longitudinal models based on logit link or survival analysis.

References

Atkinson A.C., Riani M. (2000). *Robust Diagnostic Regression Analysis*, Springer Verlag, New York.

Dobson A.J. (2002). *Introduction to Generalised Linear Model*, Chapman & Hall.

Hand D., Krzanowski W.J. (2009). *ROC curves for continuous data*, Chapman & Hall.

Contributed Session 31

Miscellanea

Spatial clustering for local analysis

Federico Benassi, Chiara Bocci, Alessandra Petrucci

The need for statistical information at detailed territorial level has greatly increased in recent years. This need is often related to the identification of spatially contiguous and homogeneous areas according to the phenomenon studied.

The aim of the paper lies in a review of methods for the analysis and detection of spatial clusters and in the application of a recently proposed clustering method. In particular, we discuss the nature and the developments of spatial data mining with special emphasis on spatial clustering and regionalization methods and techniques (Guo, 2008).

We present an original application using data from the statistical office of the city of Florence and the population census held in 2001. The first step of the analysis is devoted to describe the structure of the population of the study area. Then, we implement a regionalization model in order to get a classification of the study area into a number of homogeneous (with respect to the demographic structure) and spatially contiguous zones.

The empirical application shows that ignoring spatial clustering can lead to misleading inference and that, on the other hand, the use of appropriate methods for the detection of spatial clusters leads to meaningful inference of urban socio-economic phenomena. The results provide a considerable information to local authorities and policy makers for regional and urban planning: the application of local policies without taking into account spatial dimension can produce a lost in term of efficiency and effectiveness.

Federico Benassi, Chiara Bocci, Alessandra Petrucci,
Department of Statistics "G. Parenti", University of Florence
viale Morgagni 59 - 50134 Firenze, Italy, e-mail: alessandra.petrucci@unifi.it, benassi@ds.unifi.it

References

Guo D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning. *International Journal of Geographical Information Science 22 (7), 801-823*.

Hastie J., Tibshirani R., Friedman J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. New York.

Lloyd C. D. (2007). *Local models for spatial analysis*. Boca Renton, Florida.

Petrucci A., Brownslees C. T. (2007). Spatial Clustering Methods for the Detection of Homogenous Areas. In: *Proceedings of CLADAG 2007*, EUM, Macerata.

Symbolic tree for prognosis of localized osteosarcoma patient

Tae Rim Lee, Dae Geun Jeon, Edwin Diday

1 Introduction

Pathologic fractures occur in 5-10% of osteosarcoma patients either prior to diagnosis or during preoperative chemotherapy. Historically, the presence of a pathologic fracture was considered to contraindicate limb salvage surgery.

We hypothesized that tumor characteristics of a fractured patient rather than a fracture itself might have a prognostic influence. In order to test our hypothesis, we fit the symbolic tree model for define the features of fractured patients. In case-control study, we matched control patients by the clinical features identified in the cohort study, and analyzed the survival difference with the symbolic tree.

Using the symbolic data analysis which extend the data mining and exploratory data analysis, we can suggest the tree structured survival model on interval response variable with.

For the purpose of defining the analytic prognosis of osteosarcoma, to fit the symbolic tree structured survival model and to find the significant clinical covariates to detect the survival time from diagnosis

2 Results

The negative prognostic role of pathologic fracture in osteosarcoma is not determined, as previous case-control and retrospective cohort studies have produced con-

Tae Rim Lee,
Dept. of Information Statistics, Korea National Open University, e-mail: trlee@knou.ac.kr

Dae Geun Jeon,
Dept. of Orthopedic Surgery, Korea Cancer Center Hospital, e-mail: dgjeon@kcch.re.kr

Edwin Diday,
Statistics, University Paris Dauphine, e-mail: diday@ceremade.dauphine.fr

tradictory results. We conducted both cohort ($n = 384$) and casecontrol ($n = 111$) studies on 37 pathologically fractured localized osteosarcoma of extremity.

In cohort study, patients with a fracture showed a tendency of poorer survival, but the difference did not reach the level of significance (5-year metastasis-free survival rates; 48% for cases vs. 61% for controls; $P = 0.06$). A casecontrol study on 37 fractured and 74 control recruited from 347 patients matched for tumor size and location showed no survival difference between the cases and controls ($P = 0.12$).

References

Kim M.S., Cho W.H., Song W.S., Lee S.Y., Jeon D.G. (2007). Time dependency of prognostic factors in patients with stage II osteosarcomas. *Clinical Orthopaedics and Related Research*, 463, 157-165.

Kim M.S., Lee S.Y., Cho W.H., Song W.S., Koh J.S., Lee J.A., Yoo J.Y., Shin D.S., Jeon D.G.(2008). Growth patterns of osteosarcoma predict patient survival. *Archives of Orthopaedic and Trauma Surgery*, 129(9), 1189-96.

Mballo C., Diday E. (2005). Decision trees on interval valued variables. *The electronic journal of symbolic data analysis*, 3(1), 8-18.

Kim M.S., Lee S.Y., Lee T.R., Cho W.H., Song W.S., Koh J.S., Lee J.A., Yoo J.Y., Jeon D.G. (2009). Prognostic nomogram for predicting the 5-year probability of developing metastasis after neo-adjuvant chemotherapy and definitive surgery for AJCC stage II extremity osteosarcoma, *Annals of Oncology*, 20(5), 955-960.

Index

Agró G., 155
Arboretti Giancristofaro R., 334
Aria M., 241
Arpino B., 318

Bagnardi V., 278
Balbi S., 326
Balzanella A., 86, 230
Barbiero A., 182
Bartel H.-G., 97
Bartolucci F., 206, 345
Batagelj V., 161, 310
Behnisch M., 30
Benassi F., 366
Bennani Y., 233
Benner A., 78
Bertaccini B., 262
Bertuccelli P., 201
Bianconcini S., 337
Bigazzi R., 158
Blasius J., 329
Boari G., 222
Bocci C., 366
Bocci L., 270
Boccuzzo G., 281
Bock H.-H., 62
Bonnini S., 334
Borra S., 142
Boulesteix A.-L., 23
Bozza S., 158, 166
Brandes U., 110

Cabanes G., 233
Cadoret M., 273
Cafarelli B., 33
Cagnone S., 337
Camillo F., 46
Camiz S., 286, 350
Cantaluppi G., 222
Carissimo A., 214
Cavaliere G., 38
Cerchiello P., 321
Chirico P., 134
Chouikha H., 291
Christmann A., 43
Cipollini F., 137
Civardi M., 278
Clérot F., 89
Calabrese R., 238
Coelho Gomes G., 286
Conversano C., 65
Corsaro S., 254
Costa M., 342
Cozza V., 307
Crippa F., 278
Cutillo L., 214

D'Ambrosio A., 307
D'Esposito M. R., 257
D'Orazio M., 118
Davino C., 353
De Angelis L., 342
De Battisti F., 126
De Castris M., 174
De Iasio S., 158
De Luca A., 225
De Luca G., 153
De March D., 246
De Stefano D., 313
Deldossi L., 323
Denimal J.-J., 350
Di Bacco M., 158
di Bernardo D., 214
Di Ciaccio A., 142

Di Giuseppe E., 150
Di Salvo F., 155
Di Zio M., 118
Dias J. G., 302
Diday E., 369
Drago C., 198
Dreassi E., 81

Eichhoff M., 70
Esposito S., 150
Esposito Vinzi V., 25
Eugster M., 259

Fabbris L., 281
Facchinetti D., 265
Facchinetti S., 358
Farcomeni A., 345
Ferligoj A., 113
Ferrari D., 41
Ferrari P. A., 182
Ferretti C., 137
Fichet B., 217
Figini S., 361, 363
Friedrichs K., 73

Göker M., 102
Gantner Z., 51
Ganugi P., 137
Gattone S. A., 147
Gaul W., 49
Georgiev I., 38
Geyer-Schulz A., 275
Giordano G., 54, 115, 241
Giudici P., 185, 321, 358, 361
Giusti C., 35
Gottard A., 163
Gouzien P., 89
Grün B., 209
Greenacre M., 289
Gribov A., 94
Grossi L., 363
Grossule E., 334
Guarracino M. R., 193

Haack F., 20
Hable R., 43
Hao J.-K., 235
Heikkonen J., 305
Hornik K., 209
Hruschka H., 190

Iannario M., 57
Iodice D'Enza A., 331
Irpino A., 75, 193

Jeon D. G., 369
Jona Lasinio G., 33, 150

Kejžar N., 161
Kenett R., 169
Kestler H. A., 105
Klawonn F., 91
Korenjak-Černe S., 161
Kuhnt S., 291
Kuntz P., 235
Kurz P., 227

Lê S., 249, 273
La Vecchia D., 41
Langovaya A., 291
Lecerf F., 249
Lechevallier Y., 230
Lee T. R., 369
Leisch F., 259
Lerner J., 110, 315
Louw N., 195
Lubbers M. J., 110
Luebke K., 145
Luepke L., 59
Lursinsap C., 171

Marchetti S., 35
Marino M., 254
Marquis R., 166
Martini C., 283
Mattei A., 163
Matteucci M., 339
Mazza A., 187
McCarty C., 110
Meinfelder F., 121
Mellace G., 294
Metodiev M. V., 83
Mezzanzanica M., 137
Mignani S., 339
Misuraca M., 326
Molina J. L., 110
Monari P., 337
Montinaro M., 297
Morlini I., 219
Mucciardi M., 201
Mucha H.-J., 97

Nagel U., 110
Neri F., 46
Nissi E., 153

Ohrvik J., 243
Osmetti S. A., 238, 265, 358
Ouedraogo M., 249
Ovelgönne M., 275

Paccagnella O., 177
Pagès J., 273
Palermi S., 153
Palumbo F., 257, 331
Pandolfi S., 345
Paroli R., 323
Pasqui M., 150
Pellegrini G., 174
Pennoni F., 206
Perrotta D., 305
Petrucci A., 366
Piccolo D., 57
Pieroni L., 206
Plaia A., 155
Poli I., 246
Pollice A., 33
Polverini F., 262
Porcu M., 129
Porumbel D. C., 235
Pratesi M., 35
Punzo A., 187

Rässler S., 121
Raffinetti E., 185
Ragazzi S., 334
Ragozini G., 257, 313
Riani M., 305
Rocchetti I., 251
Rocci R., 27, 147, 294
Romano A. A., 139
Romano E., 75
Romano R., 353
Ruggieri M., 155
Röblitz S., 20

Sahmer K., 355
Salini S., 126, 169
Salmaso L., 334
Salvati N., 35
Santoro M. T., 179
Sarra A.-L., 153
Savona R., 267
Scandurra G., 139
Scanu M., 118
Scepi G., 198
Schindler D., 59
Schmidt-Thieme L., 51
Schmittbuhl M., 166
Schoier G., 243
Sciascia I., 297
Scrucca L., 99
Siciliano R., 67
Sikorski A., 227
Silachan K., 171
Solaro N., 299
Spezia L., 203
Staffieri S., 179
Stanghellini E., 347
Stefanini F. M., 107
Sulis I., 129

Tantatsanawong P., 171
Taroni F., 166
Thai-Nghe N., 51
Torti F., 305
Tschumitschew K., 91
Tutore V. A., 307

Uberti P., 361
Ultsch A., 30
Unwin A., 94

van der Putten P., 123
Vantaggi B., 347
Varriale R., 177, 211, 318
Vatolkin I., 70
Veldkamp B. P., 339
Verde R., 86, 193, 230
Vermunt J. K., 211
Vezzoli M., 267
Vignoli D., 163
Vitale M. P., 115

Weber M., 20
Weihs C., 70, 73, 145

Zani S., 219
Zargoush M., 25
Zavarrone E., 131, 326