S.Co. 2009

S.Co. 2009

Sixth Conference

COMPLEX DATA MODELING AND COMPUTATIONALLY INTENSIVE STATISTICAL METHODS FOR ESTIMATION AND PREDICTION

Politecnico di Milano, September 14-16, 2009

Proceedings

Edited by:

A.M. Paganoni, L.M. Sangalli, P. Secchi, S. Vantini

politecnica

MAGGIOLI EDITORE

ISBN 978-88387-4385-1

Maggioli Editore è un marchio di Maggioli S.p.A.
Azienda con sistema qualità certificato ISO 9001:2000

47822 Santarcangelo di Romagna (RN) • Via del Carpino, 8
Tel. 0541/628111 • Fax 0541/622020

www.maggioli.it/servizioclienti
e-mail: servizio.clienti@maggioli.it

Finito di stampare nel mese di agosto 2009
da DigitalPrint Service s.r.l., via Torricelli, 9 – Segrate (Milano)

S.Co. 2009

SIXTH CONFERENCE

COMPLEX DATA MODELING AND COMPUTATIONALLY INTENSIVE STATISTICAL METHODS FOR ESTIMATION AND PREDICTION

POLITECNICO DI MILANO, SEPTEMBER 14-16, 2009

PROCEEDINGS

Edited by:
A.M. Paganoni, L.M. Sangalli, P. Secchi, S. Vantini

MAGGIOLI EDITORE

Bayesian Analysis of D-decomposable Bidirected Graphical Models for Discrete Data

Claudia Tarantola[1] and Ioannis Ntzoufras[2]

[1] Department of Economics and Quantitative methods
University of Pavia
Via San Felice 5, 27100 Pavia, Italy (e-mail: `claudia.tarantola@unipv.it`)
[2] Department of Statistics
Athens University of Economics and Business
76 Patision Street, 10434, Athens, Greece (e-mail: `ntzoufras@aueb.gr`)

ABSTRACT. This paper deals with the Bayesian analysis of d-decomposable graphical models of marginal independence for discrete data. The model is represented by a bidirected graph with missing edges indicating marginal independence between the corresponding variables. A bidirected graphs is named d-decomposable if it is Markov equivalent to at least one DAG. We use a marginal log-linear parameterisation, under which the model is defined through suitable zero-constraints on the interaction parameters calculated within marginal distributions. We undertake a comprehensive Bayesian analysis of these models, involving suitable choices of prior distributions, estimation, model determination, as well as the allied computational issues. The methodology is illustrated with reference to a real data set.

1 Background and Preliminaries

Graphical models of marginal independence were originally introduced by Cox and Wermuth (1993) for the analysis of multivariate Gaussian distributions. They compose a family of multivariate distributions incorporating the marginal independences represented by a bidirected graph. The nodes in the graph correspond to a set of random variables and the bidirected edges represent the pairwise associations between them. A missing edge from a pair of nodes indicates that the corresponding variables are marginally independent. The list of independences implied by a bidirected graph can be obtained using the Global Markov property (Kauermann, 1996 and Richardson, 2003). The distribution of a random vector $\boldsymbol{X}_{\mathcal{V}} = \{X_v, v \in V\}$ satisfies the Global Markov property if

$$A \text{ is separated from } B \text{ by } V \setminus (A \cup B \cup C) \text{ in } G \text{ implies } X_A \perp\!\!\!\perp X_B | X_C,$$

with A, B and C disjoint subsets of V, and C may be empty. From the global Markov property, we directly derive that if two nodes i and j are disconnected then $X_i \perp\!\!\!\perp X_j$, that is the variables are marginal independent.

In the Gaussian case marginal independences correspond to zero constraints in the variance-covariance matrix. The situation is more complicated in the discrete case, where marginal independences correspond to non linear constraints on the set of parameters. Only recently parameterisations for these models have been proposed by Lupparelli *et al.* (2009) and Drton and Richardson (2008). According to Drton and Richardson (2008), a discrete graphical model of marginal independence, associated to a bidirected graph G, is a family $P(G)$ of joint distributions for a categorical random vector $\boldsymbol{X}_{\mathcal{V}}$ satisfying the global Markov property .

2 D-DECOMPOSABLE GRAPHS

Let start with some preliminary notation, for more details see e.g Drton and Richardson (2008). Let $\overline{G}$ be the skeleton of a bi-directed graph G. A triplet of nodes (v_1, v_2, v_3) in the graph $\overline{G}$ is said to be a $\vee$ configuration if (v_1, v_2) and (v_2, v_3) are edges in $\overline{G}$ while (v_1, v_3) is not. A sink orientation of a bi-directed graph G results from assigning arrows $v_1 \longrightarrow v_2 \longleftarrow v_3$ to each $\vee$ configuration (v_1, v_2, v_3) in $\overline{G}$.

A bidirected graph is d-decomposable if it is Markov equivalent to at least one DAG. Conditions under which a bidirected graph is Markov equivalent to a DAG have been stated by Pearl and Wermuth (1994) and formally proved by Drton and Richardson (2008). These results can be summarised by the following theorem

Theorem 1. *A bidirected graph G is Markov equivalent to a DAG if and only if*

(i) no edge in the sink orientation of G is bidirected;

(ii) G contains no chordless 4-chain and no chordless four cycle

If these conditions are satisfied, a Markov equivalent DAG, is constructed via an acyclic orientation of the undirected edges of its sink orientation. It worth noting that the representation in terms of a Markov equivalent DAG may not be unique. We are considering only one element of the Markov equivalent class of a DAG. In our case case this is not a problem since our model selection strategy is not affected by the selected DAG (section 4).

3 A PARAMETRISATION FOR DISCRETE BIDIRECTED GRAPH

In this paper we use the parameterisation proposed by Lupparelli *et al.* (2009) based on the following marginal log-linear model of Bergsma and Rudas (2002)

$$\boldsymbol{\lambda} = \boldsymbol{C} \log \left(\boldsymbol{M} \mathrm{vec}(\boldsymbol{\pi}) \right), \tag{1}$$

where $\boldsymbol{\pi}$ is the vector of joint probabilities, and $\mathrm{vec}(\boldsymbol{\pi})$ is a vector obtained by rearranging the elements $\boldsymbol{\pi}$ in a reverse lexicographical ordering of the corresponding variable levels with the level of the first variable changing first (or faster). The log-linear parameters $\boldsymbol{\lambda}$ are obtained from a specific set of marginal tables, and satisfies sum-to-zero constraints. With $\boldsymbol{M}$ we denote the marginalization matrix which specifies from which marginal we calculate each element of $\boldsymbol{\lambda}$. Finally $\boldsymbol{C}$ is the contrast matrix. An algorithm for constructing $\boldsymbol{C}$ and $\boldsymbol{M}$ matrices is given in the Appendix of Ntzoufras and Tarantola (2008).

Following Lupparelli *et al.* (2009) to construct a parameterisation for discrete bidirected graphs, initially we need to consider a set of a parameters $\boldsymbol{\lambda}$ derived from a hierarchical ordering (see Bergsma and Rudas, 2002) of the marginals in $DS(G) \cup V$; where $DS(G)$ is the set of all disconnected components of the bidirected graph G. Then, we must set the highest order log-linear interaction parameters derived from $DS(G)$ equal to zero.

More precisely, we should consider the following steps: i) construct a hierarchical ordering of the marginals $M_i \in DS(G)$; ii) append the marginal $M = V$ (corresponding to the full table) at the end of the list if it is not already included; iii) for every marginal table $M_i \in DS(G) \cup V$ estimate all parameters of effects in M_i that have not already estimated from the preceding marginals; iv) for every marginal table $M_i \in DS(G)$, set the highest order log-linear interaction parameter equal to zero.

4 BAYESIAN CONJUGATE ANALYSIS

In this paper we use a different approach from the one by Rudas and Bergsma (2004) and Lupparelli *et al.* (2009) in order to estimate the joint distribution $\boldsymbol{\pi}$ of a graph G. We impose the constraints implied by the graph G directly on the joint probabilities $\boldsymbol{\pi}$. We consider a multinomial sampling and we work in terms of a minimal set of probability parameters expressing marginal/conditional independences and sufficiently describing the graphical models G under investigation. By this way we can always reconstruct the joint distribution for a given graph G and then simply calculate the marginal log-linear parameters directly using (1).

More precisely, if the graph G is d-decomposable it can be expressed equivalently by a directed acyclic graph $D = D(G)$ in which the vector of joint probabilities factorizes according to the structure of the D. In particular for every cell $i \in I$ (where I is the set of all cells of the table under consideration),

$$\pi(i;D) = \prod_{v \in V} \pi_{i_v|i_{pa(v;D)}}\left(i_v|i_{pa(v;D)}\right),$$

with $pa(v;D)$ the parents set of vertex v for graph D. Hence the likelihood factorizes as

$$f(\boldsymbol{n}|\boldsymbol{\pi}^D, D) = \frac{\Gamma(N+1)}{\prod_{i \in I}\Gamma\left(n(i)+1\right)} \prod_{v \in V}\left\{\prod_{i_{cl(v;D)} \in I_{cl(v;D)}} \pi_{v|pa(v;D)}\left(i_v|i_{pa(v;D)}\right)^{n\left(i_{cl(v;D)}\right)}\right\} \tag{2}$$

where $cl(v;D) = \{v\} \cup pa(v;D)$, and $\boldsymbol{\pi}^D = \left\{\boldsymbol{\pi}_{v|pa(v;D)}; v \in V\right\}$ is the probability parameter vector for graph $D = D(G)$ with each $\boldsymbol{\pi}_{v|pa(v;D)}$ defined as $\boldsymbol{\pi}_{v|pa(v;D)} = \left\{\boldsymbol{\pi}_{v|pa(v;D)}\left(i_v|i_{pa(v;D)}\right); i_v \in I_v, i_{pa(v;D)} \in I_{pa(v;D)}\right\}$. The above factorisation can be easily adopted to get maximum likelihood estimates analytically and avoid the iterative procedure used by Rudas and Bersgma (2004) and Lupparelli *et al.* (2009).

We use conjugate priors on the appropriate probability parameters of the above parameterisation. We initially consider a Dirichlet distribution with parameters $\boldsymbol{\alpha} = \left(\alpha(i), i \in I\right)$ for the vector of the joint probabilities of the full table $\boldsymbol{\pi} = \left(\pi(i), i \in I\right)$. Hence, for the full table the prior density is given by

$$f(\boldsymbol{\pi}) = \frac{\Gamma(\alpha)}{\prod_{i \in I}\Gamma\left(\alpha(i)\right)} \prod_{i \in I} \pi(i)^{\alpha(i)-1} = f_{\mathcal{D}i}(\boldsymbol{\pi}; \boldsymbol{\alpha}) \tag{3}$$

where $f_{\mathcal{D}i}(\boldsymbol{\pi};\ \boldsymbol{\alpha})$ is the density function of the Dirichlet distribution evaluated at $\boldsymbol{\pi}$ with parameters $\boldsymbol{\alpha} = \left(\alpha(i), i \in I\right)$ and total prior precision $\alpha = \sum_{i \in I} \alpha(i)$.

For any given marginal association graph G, the model specific prior distribution is defined by the constraints imposed by $D = D(G)$. We follow the approach of Heckerman *et al.* (1995) and we assign a Dirichlet prior on the probability parameters of each node conditionally on its parents resulting in the following prior set-up

$$f\left(\boldsymbol{\pi}^D|D\right) = \prod_{v \in V}\left\{\prod_{i_{pa(v;D)} \in I_{pa(v;D)}} f_{\mathcal{D}i}\left(\boldsymbol{\pi}_{v|i_{pa(v;D)}};\ \boldsymbol{\alpha}_{v|i_{pa(v;D)}}\right)\right\} \tag{4}$$

where, for any given $i_{pa(v;D)} \in I_{pa(v;D)}$, $\boldsymbol{\pi}_{v|i_{pa(v;D)}} = \{\pi_{v|pa(v;D)}(i_v|i_{pa(v;D)}); i_v \in I_v\}$ and $\boldsymbol{\alpha}_{v|i_{pa(v;D)}} = \{\alpha_{v\cup pa(v;D)}(i_v, i_{pa(v;D)}); i_v \in I_v\} = \{\alpha_{cl(v;D)}(i_{cl(v;D)}); i_v \in I_v\}$. It is easy to show that this prior distribution factorizes in the same manner as the likelihood, in fact

$$f(\boldsymbol{\pi}^D|D) \propto \prod_{v\in V} \left\{ \prod_{i_{cl(v;D)} \in I_{cl(v;D)}} \pi_{v|pa(v;D)}(i_v|i_{pa(v;D)})^{\alpha_{cl(v;D)}(i_{cl(v;D)})-1} \right\}.$$

Since the prior is conjugate to the likelihood the posterior can be easily derived. For the saturated model, represented by the full graph, the posterior distribution is a Dirichlet distribution with parameters $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}(i) = \alpha(i) + n(i),\ i \in I)$. For the remaining models the posterior distribution is of the form in equation (4) with parameters $\widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v;D)}}$ given by $\widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v;D)}} = \left(\widetilde{\alpha}_{cl(v;D)}(i_{cl(v;D)}) = \alpha_{cl(v;D)}(i_{cl(v;D)}) + n_{cl(v;D)}(i_{cl(v;D)}),\ i_v \in I_v\right)$ for any given configuration $i_{pa(v;D)} \in I_{pa(v;D)}$.

In order to make the prior distributions 'compatible' across models, we define the prior parameters $\boldsymbol{\alpha}_{v|i_{pa(v;D)}}$ involved in the prior of $\boldsymbol{\pi}^D$ from the corresponding parameters of the prior distribution (3) for the saturated model; see Dawid and Lauritzen (2000), Roverato and Consonni (2004). Hence we specify the parameters as

$$\alpha_M(i_M) = \sum_{\{j\in I: j_M = i_M\}} \alpha(j), \tag{5}$$

for every marginal $M \subseteq V$. Compatibility is required in order to obtain consistent and coherent results across different models and within the same model under different parametrisation or factorisation. The compatible prior (4) with parameters (5) results in equal marginal likelihoods to DAGs belonging to the same Markov equivalent class. This is the so called likelihood equivalence property which in machine learning literature is frequently termed as Bayesian Dirichlet metric equivalence or BDe in short (Buntine, 1991; Heckerman *et al.*, 1995).

The prior parameters of the Dirichlet distribution are specified by using the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000). We use their approach to advocate sensible values for the Dirichlet prior parameters on the full table and the corresponding induced values for the rest of the graphs.

Let us consider an imaginary set of data represented by the frequency table $\boldsymbol{n}^* = (n^*(i), i \in I)$ of total sample size $N^* = \sum_{i\in I} n^*(i)$ and a Dirichlet 'pre-prior' with all parameters equal to α_0. Then the unnormalised prior distribution can be obtained by the product of the likelihood of $\boldsymbol{n}^*$ raised to a power w multiplied by the 'pre-prior' distribution, hence

$$\begin{aligned} f(\boldsymbol{\pi}) &\propto f(\boldsymbol{n}^*|\boldsymbol{\pi})^w \times f_{\mathcal{D}i}(\boldsymbol{\pi};\ \alpha(i) = \alpha_0, i \in I) \\ &\propto \prod_{i\in I} \pi(i)^{w n^*(i) + \alpha_0 - 1} \end{aligned}$$

For details on how to appropriately choose the values of w and $\alpha(i)$ see section 4.3 in Ntzoufras and Tarantola (2008). Finally, the posterior distribution of the marginal log-linear parameters can be obtained using straightforward Monte Carlo simulations, see Ntzoufras and Tarantola (2008) for the details.

5 ALCOHOL DATA

We now examine a well known data set presented by Knuiman and Speed (1988) regarding a small study held in Western Australia on the relationship between Alcohol intake (A), Obesity (O) and High blood pressure (H); see Table 1.

Table 1. Alcohol Data

		Alcohol intake (drinks/days)			
Obesity	High BP	0	1-2	3-5	6+
Low	Yes	5	9	8	10
	No	40	36	33	24
Average	Yes	6	9	11	14
	No	33	23	35	30
High	Yes	9	12	19	19
	No	24	25	28	29

In Table 2 we report posterior model probabilities and corresponding Log-marginal likelihoods for each models under 4 different prior set-ups. Under all prior set-ups the posterior model probability is concentrated on models H+A+O, HA+O and HO+A. Empirical Bayes and UIP-Perks' support the independence model (with posterior model probability of 0.878 and 0.807 respectively) whereas Jeffreys' and Unit Expected support a more complex structure, HO+A (with posterior model probability of 0.837 and 0.859 respectively). For the posterior summaries for model parameters see
`http://stat-athens.aueb.gr/~jbn/papers/paper21.htm`.

6 CURRENT RESEARCH

We are currently working on Bayesian analysis using non-conjugate priors on the log-linear parameters, and on the analysis of bidirected graph that do not admit a direct representation in terms of Dags. Some preliminary results will be presented during the talk.

REFERENCES

BERGSMA, W. P. AND RUDAS, T. (2002): Marginal log-linear models for categorical data. *Annals of Statistics*, *30*, 140-159.

BUNTINE, W. (1991): Theory refinement on Bayesian networks. *In* D'Ambrosio, P.S.B. and Bonissone, P. (eds.), *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 52-60. Morgan Kaufmann.

CHEN, M.H., IBRAHIMB, J.G. AND SHAO, Q. M. (2000): Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, *84*, 121-137.

Table 2. Alcohol data: Posterior model probabilities and the corresponding Log marginal likelihoods; empty cell in posterior probabilities means that it is lower than 0.0001

Posterior model probabilities (%)

	Model							
Prior Distribution	H+A+O (1)	HA+O (2)	HO+A (3)	AO+H (4)	HA+HO (5)	HA+AO (6)	HO+AO (7)	HAO (8)
Jeffreys	11.56	4.76	83.68					
Unit Expected Cell	6.91	7.21	85.88					
Empirical Bayes	87.81	0.07	12.12					
Perks	80.67	0.15	19.18					

Log-marginal likelihood for each model

	Model							
Prior Distribution	H+A+O (1)	HA+O (2)	HO+A (3)	AO+H (4)	HA+HO (5)	HA+AO (6)	HO+AO (7)	HAO (8)
Jeffreys	-79.22	-80.11	-77.24	-87.73	-90.44	-100.93	-98.06	-98.95
Unit Expected Cell	-78.51	-78.47	-75.99	-84.70	-85.27	-93.99	-91.51	-91.46
Emprirical Bayes	-86.96	-94.10	-88.94	-107.26	-124.75	-143.06	-137.91	-145.04
Perks	-86.90	-93.19	-88.33	-107.10	-121.13	-139.89	-135.03	-141.33

COX, D. R. AND WERMUTH, N. (1993): Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, *8*, 204-218, 247-277.

DAWID, A.P. AND LAURITZEN, S.L. (2000): Compatible prior distributions. In *Bayesian Methods with Applications to Science Policy and Official Statistics. The sixth world meeting of the International Society for Bayesian Analysis* (ed. E.I. George), 109-118.

DRTON, M. AND RICHARDSON, T. S. (2008): Binary models for marginal independence. *Journal of the Royal Statistical Society*, Ser. B , *70*, 287-309.

HECKERMAN, D., GEIGER, D. AND CHICKERING, D. (1995): Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*, 194-243.

IBRAHIM J.G. AND CHEN M. H. (2000): Power Prior Distributions for Regression Models. *Statistical Science*, *15*, 46-60.

KAUERMANN, G. (1996). On a dualization of graphical Gaussian models. *Scandinavian Journal of Statistics*, *23*, 105-116.

LUPPARELLI, M., MARCHETTI, G. M., BERSGMA, W. P. (2009): Parameterization and fitting of bi-directed graph models to categorical data. *Scandinavian journal of Statistics* (Forthcoming).

NTZOUFRAS, I AND TARANTOLA, C. (2008): Bayesian Analysis of Marginal Log-Linear Graphical Models for Three Way Contingency Tables. *Technical report.* arXiv:0807.1001.

PEARL, J. AND WERMUTH, N. (1994): When can association graphs admit a causal interpretation? In P. Cheesman and W. Oldford, eds., *Models and data, artifical intelligence and statistics iv.* Springer, New York, 205-214.

RICHARDSON, T. S. (2003): Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics 30*, 145-157.

ROVERATO, A. AND CONSONNI, G. (2004): Compatible Prior Distributions for DAG models. *Journal of the Royal Statistical Society*, series B, *66*, 47-61.