Computational Statistics and Data Analysis 66 (2013) 161–177

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Conjugate and conditional conjugate Bayesian analysis of discrete graphical models of marginal independence

Ioannis Ntzoufras [a,*], Claudia Tarantola [b]

[a] Department of Statistics, Athens University of Economics and Business, 76 Patision Str, 10434, Athens, Greece
[b] Department of Economics and Management, University of Pavia, Via S. Felice 5, 27100, Pavia, Italy

ARTICLE INFO

Article history:
Received 1 June 2012
Received in revised form 1 March 2013
Accepted 8 April 2013
Available online 30 April 2013

Keywords:
Bi-directed graph
Chib's marginal likelihood estimator
Contingency tables
Markov equivalent DAG over the observed margin
Monte Carlo computation

ABSTRACT

A conjugate and conditional conjugate Bayesian analysis is presented for bi-directed discrete graphical models, which are used to describe and estimate marginal associations between categorical variables. To achieve this, each bi-directed graph is re-expressed by a Markov equivalent, over the observed margin, directed acyclic graph (DAG). This DAG equivalent model is obtained using the same vertex set or with the addition of some latent variables when required. It is characterised by a minimal set of marginal and conditional probability parameters. Hence compatible priors based on products of Dirichlet distributions can be applied. For models with DAG representation on the same vertex set, the posterior distribution and the marginal likelihood is analytically available, while for the remaining ones a data augmentation scheme introducing additional latent variables is required. For the latter, the marginal likelihood is estimated using Chib's estimator. Additional implementation details including identifiability of such models are discussed. Moreover, analytic details concerning the computation of the posterior distributions of the marginal log-linear parameters are provided. The computation is achieved via a simple transformation of the simulated values of the probability parameters of the bi-directed model under study. The marginal log-linear parameterisation provides a straightforward interpretation in terms of log-odds ratios on specific marginals quantifying the associations between variables involved in the corresponding marginal. The proposed methodology is illustrated using a popular 4-way dataset.

1. Introduction

Graphical models of marginal independence were originally introduced by Cox and Wermuth (1993) for the analysis of multivariate Gaussian distributions, and subsequently extended to the discrete case by Drton and Richardson (2008a), Lupparelli (2006), Evans and Richardson (in press) and Lupparelli et al. (2009). They belong to the wider class of acyclic mixed graph models (Richardson, 2003), and they compose a family of multivariate distributions incorporating the marginal independences represented by a bi-directed graph. The vertices in the graph correspond to a set of random variables, and the edges represent the pairwise associations between them. A missing edge from a pair of vertices indicates that the corresponding variables are marginally independent.

Despite the increasing interest in the literature for graphical models of marginal independence, Bayesian analysis has not been developed as much as traditional methods. In the Gaussian case, the problem has been successfully treated by Silva and Ghahramani (2009a,b) for mixed graph models and by Khare and Rajaratnam (2011) for Gaussian decomposable covariance

* Corresponding author. Tel.: +30 210 8203968; fax: +30 210 8203968.
E-mail addresses: ntzoufras@aueb.gr (I. Ntzoufras), ctaranto@eco.unipv.it (C. Tarantola).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.04.005

graphical models. On the other hand, for the discrete case, only some initial results have been introduced; see e.g. Silva and Ghahramani (2009a), Bartolucci et al. (2012), and Ntzoufras and Tarantola (2012).

In this paper, we extend the work of Ntzoufras and Tarantola (2012) and we present a conjugate and conditional conjugate Bayesian analysis of discrete graphical models of marginal independences. We exploit the connection between bi-directed graphs and directed acyclic graphs (DAGs). A bi-directed graph can be always represented in terms of a Markov equivalent DAG over the observed margin, with the same set of vertices or some additional ones representing hidden or latent variables. A model graphically depicted by a Markov equivalent DAG, over the observed margin, describes the same independences between the observable variables as the original bi-directed graph model. The model is parameterised in terms of a minimal set of marginal and conditional probability parameters, on which we assign conjugate priors based on products of Dirichlet distributions; see Heckerman et al. (1995). Since we are working with probabilities in DAGs, the obtained parameterisation is always variation independent. The marginal likelihood for models with DAG representation including latent variables is computed using the estimator of Chib (1995). Monte Carlo simulations are used to obtain the posterior distributions of the corresponding marginal log-linear parameters which have log-odds interpretation referring to marginal dependences.

The plan of the paper is as follows. In Section 2, we introduce discrete graphical models of marginal independence, we establish the notation and we discuss their representation in terms of Markov equivalent DAGs over the observed margin. In Section 3, we present the probability parameterisation, the augmented likelihood factorisation, and the prior set-up. Section 4 is devoted to posterior inference, with particular emphasis on models with no direct DAG representation. The methodology is illustrated in Section 5 which presents the analysis of Coppen's (1966) dataset. Finally, in Section 6, we conclude with a brief discussion.

2. Discrete graphical models of marginal independence

2.1. Bi-directed graphs and Markov properties

In this section we briefly introduce graphical models of marginal independence, the related notation and terminology; for more details see, for example, in Drton and Richardson (2008a).

A bi-directed graph $G = (\mathcal{V}, E)$, is a graph with vertex set $\mathcal{V}$, and edge set E, such that $(v_\imath, v_\jmath) \in E$ if and only if $(v_\jmath, v_\imath) \in E$. We denote each bi-directed edge by $(\overleftrightarrow{v_i, v_j}) = \{(v_i, v_j), (v_j, v_i)\}$ and, following Richardson (2003), we represent it with a bi-directed arrow. An alternative representation, proposed by Cox and Wermuth (1993), is by undirected dashed edges.

The skeleton $\overline{G}$ of a bi-directed graph G is the graph obtained by making all edges undirected; every triplet of vertices $(v_\imath, v_\jmath, v_k)$ in $\overline{G}$ with edges $(v_\imath, v_\jmath)$ and $(v_\jmath, v_k)$ and with no edge connecting v_i and v_k is named $\vee$ configuration. A path connecting two vertices, v_0 and v_m, is a finite sequence of distinct vertices $v_0, \dots, v_m$ such that $(v_{i-1}, v_\imath), i = 1, \dots, m$, is an edge of the graph. A vertex set $\mathcal{C} \subseteq \mathcal{V}$ is connected if every two vertices $v_\imath$ and $v_\jmath$ are joined by a path in which every vertex is in $\mathcal{C}$. The vertex set $\mathcal{C} \subseteq \mathcal{V}$ induces a subgraph $G_\mathcal{C}$ obtained keeping only the edges having both end points in $\mathcal{C}$.

The graph is used to represent marginal independences between a set of discrete random variables $X_\mathcal{V} = (X_v, v \in \mathcal{V})$, each one taking values $i_v \in \mathcal{I}_v$; where $\mathcal{I}_v$ is the set of possible levels for variable v. The cross-tabulation of variables $X_\mathcal{V}$ produces a $|\mathcal{V}|$-way contingency table with cell frequencies $\boldsymbol{n} = (n(i), i \in \mathcal{I})$ where $\mathcal{I} = \times_{v \in \mathcal{V}} \mathcal{I}_v$. Similarly for any marginal $M \subseteq \mathcal{V}$, we denote with $X_M = (X_v, v \in M)$ the set of variables which produce the marginal table with frequencies $\boldsymbol{n}_M = (n_M(i_M), i_M \in \mathcal{I}_M)$ where $\mathcal{I}_M = \times_{v \in M} \mathcal{I}_v$.

The list of independences implied by a bi-directed graph can be obtained using the following Markov properties: the *pairwise Markov* property (Cox and Wermuth, 1993) and the *connected set Markov* property (Richardson, 2003). The distribution of a random vector $X_\mathcal{V}$ satisfies the pairwise Markov property, if a missing edge in the graph indicates marginal independence between the corresponding variables. The distribution of a random vector $X_\mathcal{V}$ satisfies the connected set Markov property if for every disconnected set $\mathcal{D}$ the subvectors $X_{\mathcal{C}_1}, X_{\mathcal{C}_2}, \dots, X_{\mathcal{C}_r}$, corresponding to its connected components $\mathcal{C}_1, \dots, \mathcal{C}_r$, are mutually independent. For discrete variables the connected set Markov property implies the pairwise Markov property, whereas the converse is not generally true. Following Drton and Richardson (2008a), we define a discrete graphical model of marginal independence as the family of probability distributions for $X_\mathcal{V}$ that satisfy the connected set Markov property.

2.2. Representation in terms of Markov equivalent DAGs over the observed margin

A bi-directed graph can be represented via a DAG that is Markov equivalent over the observed margin either with the same vertex set (homogeneous bi-directed graph), or with the introduction of some additional latent vertices; see Pearl and Wermuth (1994), Drton and Richardson (2008b) and Silva and Ghahramani (2009a,b).

Following Letac and Massam (2007, Definition 2.1), a bi-directed graph is named homogeneous if it does not include a bi-directed 4-chain or a chordless 4-cycle as sub-graphs. Given the skeleton $\overline{G}$ of the examined graph, one should assign arrows $v_\imath \longrightarrow v_\jmath \longleftarrow v_k$ to each $\vee$ configuration $(v_\imath, v_\jmath, v_k)$ in $\overline{G}$, constructing in this way the sink orientation of G. If no edge in the sink orientation is bi-directed the graph is homogeneous. If the sink orientation contains bi-directed edges, a Markov equivalent DAG over the observed margin can be constructed substituting every bi-directed edge $v_1 \longleftrightarrow v_2$ with the directed configuration $v_1 \longleftarrow \ell \longrightarrow v_2$, where vertex ℓ represents a hidden or latent variable; see Theorem 3 in Pearl and Wermuth (1994). We then obtain a new graphical structure, with ℓ being the parent vertex of the children v_1 and v_2. Finally, a Markov equivalent DAG over the observed margin is constructed via an acyclic orientation of the undirected edges present in the sink orientation of the graph.

Any DAG which encodes the same independences between the observable variables as G will be called augmented DAG of G. More precisely, let $\mathcal{L}$ be the set of hidden or latent vertices introduced in the graph to obtain a Markov equivalent DAG over the observed margin, and $X_{\mathcal{L}} = (X_\ell, \ell \in \mathcal{L})$ be the corresponding vector of variables. The augmented DAG of G is the graph representing the relation between the variables of $X_{\mathcal{A}}$, with $\mathcal{A} = \mathcal{V} \cup \mathcal{L}$. Naturally, if the bi-directed graph is homogeneous $\mathcal{A} = \mathcal{V}$ since $\mathcal{L} = \emptyset$.

3. Model set-up

In the following, we work in terms of the augmented DAG representation of the model, parameterising it via a minimal set of marginal and conditional probability parameters sufficient to obtain the joint distribution of interest.

3.1. Probability parameterisation and augmented likelihood factorisation

Given an augmented DAG representation D, the vector of joint probabilities $\boldsymbol{p}^{\mathcal{A}}(D)$ corresponding to the augmented set of variables $X_{\mathcal{A}}$ factorises as

$$p^{\mathcal{A}}(i; D) = \prod_{v \in \mathcal{A}} \pi_{v|pa(v;D)} \left(i_v | i_{pa(v;D)} \right), \tag{1}$$

where $pa(v; D)$ stands for the parents set of vertex v in graph D, and $\pi_{v|U}(i_v|i_U)$ is the parameter for the conditional probability $p(X_v = i_v|X_U = i_U)$. The corresponding joint probabilities $\boldsymbol{p}(D) = (p(i; D), i \in \mathcal{I}_{\mathcal{V}})$ associated to the observable variables $X_{\mathcal{V}}$ are a function of $\boldsymbol{p}^{\mathcal{A}}(D)$, and are given by

$$p(i; D) = \sum_{i_\ell \in \mathcal{I}_{\mathcal{L}}} p^{\mathcal{A}}(i, i_\ell; D). \tag{2}$$

Since we focus on a specific augmented DAG, we simplify the notation by eliminating D from $pa(v; D)$, $\pi_{v|pa(v;D)} \left(i_v | i_{pa(v;D)} \right)$, $p^{\mathcal{A}}(i\,; D)$ and $p(i; D)$ appearing in (1) and (2). In the following, we work with a minimal set of probability parameters

$$\boldsymbol{\pi}^D = \left(\boldsymbol{\pi}_{v|i_{pa(v)}}; v \in \mathcal{A}, i_{pa(v)} \in \mathcal{I}_{pa(v)} \right), \quad \text{with } \boldsymbol{\pi}_{v|i_{pa(v)}} = \left(\pi_{v|pa(v)}(i_v|i_{pa(v)}); i_v \in \mathcal{I}_v \right).$$

This set refers to conditional and marginal probability parameters which are sufficient to reconstruct the joint probabilities $\boldsymbol{p}^{\mathcal{A}}$ under dependences and independences induced by D. The augmented likelihood for a specific D, is given by

$$f(\boldsymbol{n}^{\mathcal{A}}|\boldsymbol{\pi}^D) = \frac{\Gamma(N+1)}{\prod_{i_{\mathcal{A}} \in \mathcal{I}_{\mathcal{A}}} \Gamma\left(n^{\mathcal{A}}(i_{\mathcal{A}})+1\right)} \prod_{v \in \mathcal{A}} \left\{ \prod_{i_{cl(v)} \in \mathcal{I}_{cl(v)}} \pi_{v|pa(v)} \left(i_v | i_{pa(v)} \right)^{n^{\mathcal{A}}\left(i_{cl(v)}\right)} \right\}, \tag{3}$$

where $cl(v) = \{v\} \cup pa(v)$ and $\boldsymbol{n}^{\mathcal{A}} = \left(n^{\mathcal{A}}(i_{\mathcal{A}}), i_{\mathcal{A}} \in \mathcal{I}_{\mathcal{A}} \right)$ are the cell frequencies of the augmented contingency table for variables $\mathcal{A}$. If the bi-directed graph is homogeneous, the DAG representation of G does not includes any latent variables (i.e. $\mathcal{L} = \emptyset$ and $\mathcal{A} = \mathcal{V}$), hence the model likelihood is directly given by (3).

Even if every bi-directed graph can be graphically depicted via an augmented DAG, the model obtained from the augmented DAG via marginalisation may be not directly equivalent to the original one. In fact, when the augmented DAG includes latent variables, the structure of the DAG may impose additional inequality constraints on the distribution of the observed variables. The simplest case is the 4-chain bi-directed graph that can be represented via an augmented DAG containing one additional latent variable. In this case the marginal distribution of the observable should satisfy Bell's inequalities; see, for example, Tian and Pearl (2002), Ver Steeg and Galstyan (2011), Evans (2012) and Kang and Tian (2012). Hence, the augmented DAG can be considered as a close approximation of the corresponding marginal association model unless all inequalities become inactive a posteriori (i.e. are satisfied by the posterior distribution of the probabilities of the marginal association model).

3.2. Prior distributions

We use conjugate priors based on products of Dirichlet distributions; see Heckerman et al. (1995). We assign a Dirichlet prior on the probability parameters of each vertex conditionally on its parents resulting in the following prior set-up

$$\begin{aligned} f\left(\boldsymbol{\pi}^D\right) &= \prod_{v \in \mathcal{A}} \prod_{i_{pa(v)} \in \mathcal{I}_{pa(v)}} f_{\mathcal{D}i}\left(\boldsymbol{\pi}_{v|i_{pa(v)}}; \boldsymbol{\alpha}_{v|i_{pa(v)}} \right) \\ &\propto \prod_{v \in \mathcal{A}} \left\{ \prod_{i_{cl(v)} \in \mathcal{I}_{cl(v)}} \pi_{v|pa(v)} \left(i_v | i_{pa(v)} \right)^{\alpha_{cl(v)}\left(i_{cl(v)}\right)-1} \right\}, \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}_{v|i_{pa(v)}} = \left(\alpha_{cl(v)}\left(i_{cl(v)}\right); i_v \in \mathcal{I}_v \right)$ and $f_{\mathcal{D}i}(\boldsymbol{\pi}; \boldsymbol{\alpha})$ is the Dirichlet density function with parameters $\boldsymbol{\alpha}$ evaluated at $\boldsymbol{\pi}$.

In order to make the prior distributions "compatible" across models, we assign a Dirichlet distribution on the vector of joint probabilities $\boldsymbol{p}$ for the saturated model of the observed table with parameters $\boldsymbol{\alpha} = (\alpha(i), i \in \mathcal{I})$. If the model is nonhomogeneous, we use a similar Dirichlet distribution with parameters $\alpha^{\mathcal{A}}(i)$ (for all $i \in \mathcal{I}^{\mathcal{A}}$) for the vector of joint probabilities $\boldsymbol{p}^{\mathcal{A}}$ of the saturated model on the augmented contingency table, such that the prior on $\boldsymbol{p}$ is the same as the one considered initially. Thus, we obtain compatibility by setting $\alpha(i) = \sum_{i_\ell \in \mathcal{I}_\mathcal{L}} \alpha^{\mathcal{A}}(i, i_\ell)$. Under this approach, each component $\boldsymbol{\pi}_{v|i_{pa(v)}}$ of $\boldsymbol{\pi}^D$ will a priori follow the Dirichlet distribution appearing in (4) with each parameter calculated as

$$\alpha_{cl(v)}\left(i_{cl(v)}\right) = \sum_{i_{\mathcal{A}\setminus cl(v)} \in \mathcal{I}_{\mathcal{A}\setminus cl(v)}} \alpha^{\mathcal{A}}\left(i_{cl(v)}, i_{\mathcal{A}\setminus cl(v)}\right).$$

More details on compatible prior distributions can be found in Dawid and Lauritzen (2000), Roverato and Consonni (2004), and Consonni and Veronese (2008).

When no prior information is available, a usual choice is to consider equal $\alpha(i)$ for all cells $i \in \mathcal{I}$. Common choices for $\alpha(i)$ are 1/2 (Jeffreys prior), 1 (unit expected cell prior, UEC) and $1/|\mathcal{I}|$ (Perks, 1947); see Dellaportas and Forster (1999) for additional details. The choice of this prior parameter value is of prominent importance for the model comparison due to the well known sensitivity of the posterior model odds and the Bartlett–Lindley paradox Bartlett (1957) and Lindley (1957). In this paper, we present results for the previous prior choices. A detailed comparison of prior choices is presented in Table 2 of Ntzoufras and Tarantola (2012).

4. Posterior inference

4.1. Conditional conjugate analysis for nonhomogeneous models

From (3) and (4), the posterior distribution of the parameters $\boldsymbol{\pi}^D$ given a set of augmented data $\boldsymbol{n}^{\mathcal{A}}$ is given by

$$f(\boldsymbol{\pi}^D|\boldsymbol{n}^{\mathcal{A}}) = \prod_{v\in\mathcal{A}} \prod_{i_{pa(v)}\in\mathcal{I}_{pa(v)}} f_{\mathcal{D}i}\left(\boldsymbol{\pi}_{v|i_{pa(v)}}; \widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v)}}\right), \tag{5}$$

where $\widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v)}} = \left(\widetilde{\alpha}_{cl(v)}(i_{cl(v)}) = n^{\mathcal{A}}_{cl(v)}(i_{cl(v)}) + \alpha_{cl(v)}(i_{cl(v)}), i_v \in \mathcal{I}_v\right)$, for any given configuration $i_{pa(v)} \in \mathcal{I}_{pa(v)}$.

Moreover, the posterior distribution of the frequencies of the augmented table, $f(\boldsymbol{n}^{\mathcal{A}}|\boldsymbol{n}, \boldsymbol{\pi}^D)$, is given by

$$\begin{aligned} f(\boldsymbol{n}^{\mathcal{A}}|\boldsymbol{n}, \boldsymbol{\pi}^D) &\propto \prod_{i\in\mathcal{I}_\mathcal{V}} \prod_{i_\ell\in\mathcal{I}_\mathcal{L}} p^{\mathcal{A}}(i, i_\ell)^{n^{\mathcal{A}}(i,i_\ell)} I\left(n(i) = n^{\mathcal{A}}_{\mathcal{V}}(i)\right) \\ &= \prod_{i\in\mathcal{I}_\mathcal{V}} f_m\left(\boldsymbol{n}^{\mathcal{A}}_{\mathcal{V}}(i, \bullet); \boldsymbol{\varpi}(i), n(i)\right), \end{aligned} \tag{6}$$

with

$$\boldsymbol{\varpi}(i) = \left(\varpi(i) = \frac{p^{\mathcal{A}}(i, i_\ell)}{\sum_{i_L\in\mathcal{I}_\mathcal{L}} p^{\mathcal{A}}(i, i_L)}; i_\ell \in \mathcal{I}_\mathcal{L}\right), \tag{7}$$

$\boldsymbol{n}^{\mathcal{A}}_{\mathcal{V}}(i, \bullet)$ being the $|\mathcal{I}_\mathcal{L}|$ dimensional vector of cell frequencies with elements $\{n^{\mathcal{A}}(i, i_\ell)$ for all $i_\ell \in \mathcal{I}_\mathcal{L}\}$, and $n^{\mathcal{A}}_{\mathcal{V}}(i) = \sum_{i_\ell\in\mathcal{I}_\mathcal{L}} n^{\mathcal{A}}(i, i_\ell)$, for any given $i \in \mathcal{I}$. Moreover, we denote by $I(\cdot)$ the indicator function and by $f_m\,(\boldsymbol{n}; \boldsymbol{\pi}, N)$ the probability function of the multinomial distribution with probability vector $\boldsymbol{\pi}$ and N independent trials evaluated at $\boldsymbol{n}$.

In order to obtain a sample from the posterior distribution of $\boldsymbol{\pi}^D$ we consider the following Gibbs algorithm generating sequentially values from (5) and (6):

(i) Generate the frequencies $\boldsymbol{n}^{\mathcal{A}}$ of an augmented table by randomly splitting every single cell $n(i)$ using $\boldsymbol{n}^{\mathcal{A}}_{\mathcal{V}}(i, \bullet) \sim \mathcal{M}ultinomial\,(\boldsymbol{\varpi}(i), n(i))$, for every $i \in \mathcal{I}$.
(ii) For every $v \in \mathcal{A}$ and $i_{pa(v)} \in \mathcal{I}_{pa(v)}$ generate $\boldsymbol{\pi}_{v|i_{pa(v)}} \sim \mathcal{D}irichlet\left(\widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v)}}\right)$.

The second step of the algorithm should be applied only to parameters without any identifiability constraints; see Section 4.2 for more details. If one or more parameters in vector $\boldsymbol{\pi}_{v|i_{pa(v)}}$ are constrained, then the corresponding conditional Dirichlet distribution must be used in the conditional posterior distribution; for the properties of the Dirichlet distribution see, for example, in Table 2 of Frigyik et al. (2010).

Finally, for both homogeneous and nonhomogeneous models, we can easily obtain a sample from the posterior distribution of the joint probabilities $\boldsymbol{p}$ by simply transforming each observation of the simulated sample of $\boldsymbol{\pi}^D$ using (1) and then summing over all levels of the latent variables as given by (2).

4.2. Some important implementation details

The use of DAGs with latent variables to represent nonhomogeneous models creates two problems which are common in latent variable modelling: non-identifiability and label switching.

Let us consider the identifiability problem first. In order to remain in the class of Markov equivalent DAGs (over the observed margin), the number of levels of the introduced latent variables should be such that the augmented DAG model has at least as many parameters as the one represented by the original bi-directed graph. If the dimension is smaller, then the new model will impose additional undesirable dependences or other constraints that are not implied by the original bi-directed graph. This can be possibly avoided if the dimension is larger, but still the new augmented model should be handled with caution since other constraints may change the original model structure.

Having this in mind, we suggest the following rules of thumb. First, we introduce latent variables, with the least possible number of levels, satisfying the restriction that the augmented DAG should have at least the same number of parameters as the original bi-directed model. This can be found heuristically by initially introducing binary latent variables and continue by increasing the number of levels until the previous restriction is satisfied. At the second stage, we impose a number of constraints equal to the difference between the number of parameters of the augmented DAG and the corresponding bi-directed model. We suggest to first impose constraints on the probabilities of the latent variables and then continue, if necessary, to the probability parameters of the first level of each child in D with at least one latent parent conditioned on the first levels of its parents. We propose to set the constrained parameters equal to the prior mean we would place on the corresponding parameters of the unconstrained version of the model.

Thus, starting from the probabilities of the latent variables we set $\pi_\ell\,(i_\ell) = \alpha_\ell\,(i_\ell)/\sum_{i_\ell \in \mathcal{I}_\ell} \alpha_\ell\,(i_\ell)$ and

$$\pi_{v|pa(v)}\left(i_v = 1 | i_{pa(v)} = \{1\}^{|pa(v)|}\right) = \frac{\alpha_{cl(v)}\left(i_v = 1, i_{pa(v)} = \{1\}^{|pa(v)|}\right)}{\sum\limits_{i_v \in \mathcal{I}_v} \alpha_{cl(v)}\left(i_v = 1, i_{pa(v)} = \{1\}^{|pa(v)|}\right)},$$

for specific $v \in \mathcal{V}$ and its parents. For prior distributions with equal $\alpha(i)$ (as the prior set-ups we use here), these constraints simplify to $\pi_\ell(i_\ell) = 1/|\mathcal{I}_\ell|$ for $\ell \in \mathcal{L}$ and $\pi_{v|pa(v)}\left(i_v = 1|i_{pa(v)} = \{1\}^{|pa(v)|}\right) = 1/|\mathcal{I}_v|$ for specific $v \in \mathcal{V}$ and its parents. This is indeed the parameterisation we have used in the illustration of Section 5. Note that if one or more parameters in a vector $\boldsymbol{\pi}_{v|i_{pa(v)}}$ are constrained, then the prior distribution (4) should be modified using the corresponding conditional Dirichlet distributions.

An alternative is to implement the MCMC algorithm described in Section 4.1 on the unconstrained model. When informative priors are used, then constraints are indirectly imposed by them and the MCMC will produce results from the posterior distribution without any problem (returning the prior as posterior for unidentifiable variables). If flat, non-informative prior distributions are used, the MCMC output for the model parameters $\boldsymbol{\pi}^D$ will present a non-convergence picture. Nevertheless, both joint probabilities $\boldsymbol{p}$ and marginal log-linear parameters $\boldsymbol{\lambda}$ will converge to the appropriate target posterior distributions since they are both well defined. Therefore, a possible solution is to leave the MCMC run on the unconstrained model and focus on the interpretation of $\boldsymbol{p}$ and $\boldsymbol{\lambda}$.

Concerning the label switching problem, many approaches have been proposed in the literature such as imposing inequality constraints (see, e.g., Diebolt and Robert (1994)), re-labelling algorithms (see, e.g., Stephens (2000)), the random permutation sampler of Frühwirth-Schnatter (2001), and many others (see, e.g., Papastamoulis and Iliopoulos (2010)); see in Jasra et al. (2005) and Yao (2012) for a nice overview of the subject. Nevertheless, for the bi-directed 4-chain graphs we have implemented, the MCMC was exploring only one of the alternative modes and therefore not causing any problems in the posterior inference. Even for bi-directed chordless 4-cycle graphs, where label switching is more intense due to the multiple permutations of the latent configurations, the joint probabilities and the log-linear parameters are not affected by this behaviour since they are identifiable with unimodal posterior distributions. To confirm this, we have performed extensive experimental tests with different starting points and simulated datasets. In all cases both the joint probabilities and the log-linear parameters (Figs. 4 and 7) were quite robust and comparable with MLEs and with other MCMC methods implemented directly on log-linear models (work currently in progress by the authors). Therefore, we have not pursued this issue further except for the computation of the marginal likelihood estimate where a correction for the label switching problem was implemented as we describe in Section 4.4.

Finally, for bi-directed 4-chain graphs, we can check whether Bell's inequalities are active by monitoring the posterior distribution of the expression involved in them. This can be easily done using the MCMC output; for each iteration we can directly calculate the corresponding value of the expression involved in each Bell's inequality. In the case that this posterior distribution is uni-modal, and the inequality is actively involved in the posterior inference, then the posterior mode will be close to the imposed bound. If it is not actively involved in the posterior inference, then the posterior distribution of this expression will be well placed away from the imposed bound; see Section 5 for an illustration.

4.3. Marginal log-linear parameters estimation

Alternative parameterisations for discrete graphical models of marginal independence have been proposed in the literature by Lupparelli (2006), Drton and Richardson (2008b), Lupparelli et al. (2009), Evans and Richardson (in press) and Roverato et al. (2013). In particular, Lupparelli (2006) and Lupparelli et al. (2009) introduced the use of a marginal log-linear parameterisation in bi-directed graph models. This parameterisation has the advantage that each parameter is a log-odds ratio measuring local marginal dependences between variables. On the other hand, the probability parameters of the Markov equivalent DAG (over the observed margin) are variation independent which is not always true for the alternative marginal log-linear parameters (see, for example, Lupparelli et al. (2009), and Evans and Richardson (in press)) causing important problems and inconsistencies in the implementation of MCMC methods. Moreover, under the DAG-probability parameterisation, the likelihood is directly available while for the log-linear parameters an iterative procedure is required for the calculation of the joint probabilities. In this section, we introduce the marginal log-linear parameterisation for bi-directed graphs and describe how the corresponding posterior estimates can be obtained from our approach.

The marginal log-linear parameterisation for bi-directed graphs is based on the class of marginal log-linear models of Bergsma and Rudas (2002). The marginal log-linear parameters can be obtained by

$$\boldsymbol{\lambda} = \boldsymbol{C} \log \left(\boldsymbol{M} \mathrm{vec}(\boldsymbol{p}) \right), \tag{8}$$

where $\mathrm{vec}(\boldsymbol{p})$ is a vector of dimension $|\mathcal{I}|$ obtained by rearranging the elements $\boldsymbol{p}$ in a reverse lexicographical ordering of the corresponding variable levels with the level of the first variable changing first (or faster). The parameter vector $\boldsymbol{\lambda}$ satisfies sum-to-zero constraints, and $\boldsymbol{C}$ indicates the corresponding contrast matrix. Finally $\boldsymbol{M}$ is the marginalisation matrix which specifies from which marginal we calculate each element of $\boldsymbol{\lambda}$. Details for the construction of $\boldsymbol{C}$ and $\boldsymbol{M}$ are available in the Appendix of Ntzoufras and Tarantola (2012).

In order to obtain a marginal log-linear parameterisation for a bi-directed graph G, the disconnected sets of the graph should be considered as marginals, with eventually the addition of the full set of variables (if the graph is connected). The marginal should be arranged according to a hierarchical ordering (see Bergsma and Rudas (2002)). Then zero constraints on specific marginal log-linear parameters are imposed; see Lupparelli et al. (2009) for more details.

The marginal log-linear modelling set-up is expressed in terms of log-odds ratios referring to specific marginal tables. When an edge is absent from the bi-directed graph G, then the corresponding λ parameters (i.e. the corresponding log-odds ratio) are constrained to zero. This parameterisation is useful in cases when information is available for specific marginal associations via odds ratios (i.e. marginal log-linear parameters) or when partial information (i.e. marginals) is available. Unfortunately Eq. (8) cannot be used to obtain a closed form expression for $\boldsymbol{p}$, hence iterative procedures are needed to obtain the likelihood of the model for each set of $\boldsymbol{\lambda}$ values; see Rudas and Bergsma (2004) and Lupparelli (2006).

Working directly on graphs with marginal log-linear parameterisation is complicated. First of all, no conjugate or conditional conjugate analysis is feasible. Moreover, the likelihood cannot be written directly in a closed form. Nevertheless, with the approach presented in this work, we can estimate the posterior distribution of $\boldsymbol{\lambda}$ in a straightforward manner using Monte Carlo samples from the posterior distribution of $\boldsymbol{\pi}^D$. Specifically, a sample from the posterior distribution of $\boldsymbol{\lambda}$ can be generated by the following iterative procedure. At each iteration t (for $t = 1, \ldots, T$):

(i) Generate a random value $\boldsymbol{\pi}^{D(t)}$ from the posterior distribution of $\boldsymbol{\pi}^D$.
(ii) Calculate the full table of probabilities $\boldsymbol{p}^{(t)}$ from $\boldsymbol{\pi}^{D(t)}$.
(iii) Obtain the vector of marginal log-linear parameters, $\boldsymbol{\lambda}^{(t)}$ from $\boldsymbol{p}^{(t)}$ via Eq. (8).

The generated values $\left(\boldsymbol{\lambda}^{(t)}; t = 1, 2, \ldots, T\right)$ can be used to estimate summaries of the posterior distribution $f(\boldsymbol{\lambda}|G)$ or obtain plots fully describing this distribution. A major advantage of this approach is that all zero constraints on $\boldsymbol{\lambda}$ are automatically imposed by construction.

4.4. Chib's marginal likelihood estimator of marginal likelihood

In this Section, we illustrate how Chib's (1995) estimator can be used to evaluate the marginal likelihood for nonhomogeneous models. An estimate of the marginal likelihood is given by

$$\widehat{f}(\boldsymbol{n}|D) = \frac{f(\boldsymbol{n}|\boldsymbol{\pi}^{*D}) f(\boldsymbol{\pi}^{*D})}{f(\boldsymbol{\pi}^{*D}|\boldsymbol{n})}, \tag{9}$$

where $\boldsymbol{\pi}^{*D}$ should be a point of high posterior density in order to get reliable estimates. The posterior mode, the posterior median or the posterior mean can be appropriate points that can be used in (9).

The likelihood is given by the probability function of a multinomial distribution with joint probabilities $\boldsymbol{p}^*$ evaluated at the observed data $\boldsymbol{n}$

$$\log f(\boldsymbol{n}|\boldsymbol{p}^*, D) = \log \Gamma \left(\sum_{i \in \mathcal{I}} n(i) + 1 \right) - \sum_{i \in \mathcal{I}} \log \Gamma \left(n(i) + 1 \right) + \sum_{i \in \mathcal{I}} n(i) \log p^*(i),$$

where $\boldsymbol{p}^*$ is given by (2) after calculating (1) with $\boldsymbol{\pi}^D = \boldsymbol{\pi}^{*D}$.

The posterior ordinate $f(\boldsymbol{\pi}^{*D}|\boldsymbol{n})$ is given by

$$f(\boldsymbol{\pi}^{*D}|\boldsymbol{n}) = \mathsf{E}_{\boldsymbol{n}^{\mathcal{A}}|\boldsymbol{n}} \left[\prod_{v \in \mathcal{A}} \prod_{i_{pa(v)} \in \mathcal{I}_{pa(v)}} f_{\mathcal{D}i} \left(\boldsymbol{\pi}^*_{v|i_{pa(v)}}; \widetilde{\boldsymbol{\alpha}}_{v|i_{pa(v)}} \right) \right],$$

where the expectations are taken with respect to the posterior distribution of the latent data $\boldsymbol{n}^{\mathcal{A}}$. The above equation results directly from the procedure described in Section 2.1.2 of Chib (1995) by further assuming independence between $\boldsymbol{\pi}_{v|i_{pa(v)}}$

Table 1
Coppen's (1966) dataset on symptoms of psychiatric patients.

B	D	$C = 1$		$C = 2$	
		$A = 1$	$A = 2$	$A = 1$	$A = 2$
1	1	15	23	25	14
	2	9	14	46	47
2	1	30	22	22	8
	2	32	16	27	14

A: *stability*; *B*: *validity*; *C*: *acute depression*; *D*: *solidity*.

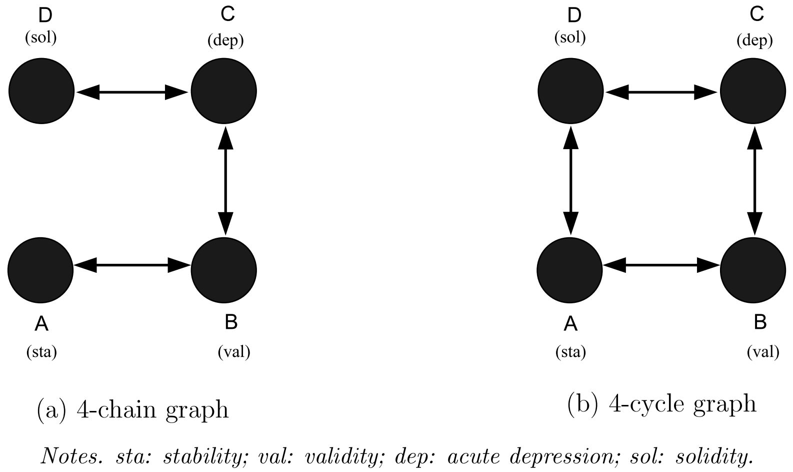


Fig. 1. Bi-directed 4-chain and chordless 4-cycle graphs fitted in Coppen's data.

given the augmented table $\boldsymbol{n}^{\mathcal{A}}$ for all $v \in \mathcal{A}$ and $i_{pa(v)} \in \mathcal{I}_{pa(v)}$ when $\boldsymbol{n}^{\mathcal{A}}$ is available. So $\widehat{f}(\boldsymbol{\pi}^{*D}|\boldsymbol{n})$ is finally estimated via

$$\widehat{f}(\boldsymbol{\pi}^{*D}|\boldsymbol{n}) = \frac{1}{T}\sum_{t=1}^{T}\left\{\prod_{v\in\mathcal{A}}\prod_{i_{pa(v)}\in\mathcal{I}_{pa(v)}}\left[f_{\mathcal{D}i}\left(\boldsymbol{\pi}^{*}_{v|i_{pa(v)}};\widetilde{\boldsymbol{\alpha}}^{(t)}_{v|i_{pa(v)}}\right)\right]\right\},$$

where $\widetilde{\boldsymbol{\alpha}}^{(t)}_{v|i_{pa(v)}} = \left(\widetilde{\alpha}^{(t)}_{cl(v)}(i_{cl(v)}) = n^{\mathcal{A}(t)}_{cl(v)}(i_{cl(v)}) + \alpha_{cl(v)}(i_{cl(v)}),\, i_v \in \mathcal{I}_v\right)$, for any given configuration $i_{pa(v)} \in \mathcal{I}_{pa(v)}$. In the above expression, the Dirichlet densities must be replaced by the corresponding conditional Dirichlet when one or more components of $\boldsymbol{\pi}_{v|i_{pa(v)}}$ are constrained. Additional details for the 4-chain and chordless 4-cycle bi-directed graphs are provided in Section 5 and in the Appendix.

Due to the label switching problem, we adjust the original estimator by the correction originally proposed by Neal (1998) and further developed in more detail by Marin and Robert (2008). Moreover, the mode (or values close to it) is most suitable choice for $\boldsymbol{\pi}^{*D}$ that can be used in the Chib's estimator since the mean and the median will be away from points of high posterior density if the MCMC explores all local modes. In cases that the MCMC visits only one of the permutations of the labels of the latent variables, then using the posterior mean and median in Chib's estimator also results in good estimates of the marginal likelihood.

5. Illustrative example: Coppen's dataset

We consider a dataset presented by Coppen (1966) regarding the interrelation between symptoms manifested by 362 psychiatric patients; see Table 1. The symptoms are: $A \equiv$ stability $u(1 =$ extroverted, $2 =$ introverted); $B \equiv$ validity ($1 =$ energetic, $2 =$ psychasthenic); $C \equiv$ acute depression ($1 =$ yes, $2 =$ no); $D \equiv$ solidity ($1 =$ hysteric, $2 =$ rigid).

This dataset has been already analysed with different type of graphical models by Wermuth (1976), Lupparelli et al. (2009) and Roverato et al. (2013). In particular Lupparelli et al. (2009) applied discrete graphical models of marginal independence to the graph in Fig. 1(a).

We present posterior results for the bi-directed 4-chain graph of Fig. 1(a) implemented by Lupparelli et al. (2009) and the closest bi-directed chordless 4-cycle graph depicted in Fig. 1(b). Moreover, we present marginal likelihoods and posterior probabilities for all 4-vertex graphs. Posterior analysis for homogeneous models can be implemented following the procedures described in Ntzoufras and Tarantola (2012). All the analysis was performed using three prior choices for the saturated model of the observed table: the Perks prior (with $\alpha(i) = 1/2^4$), the Jeffreys prior (with $\alpha(i) = 1/2$) and the unit expected cells prior (with $\alpha(i) = 1$). The priors of all other models have been designed to be compatible with these three baseline priors following the procedure described in Section 3.2. All results were obtained using R version 2.12.

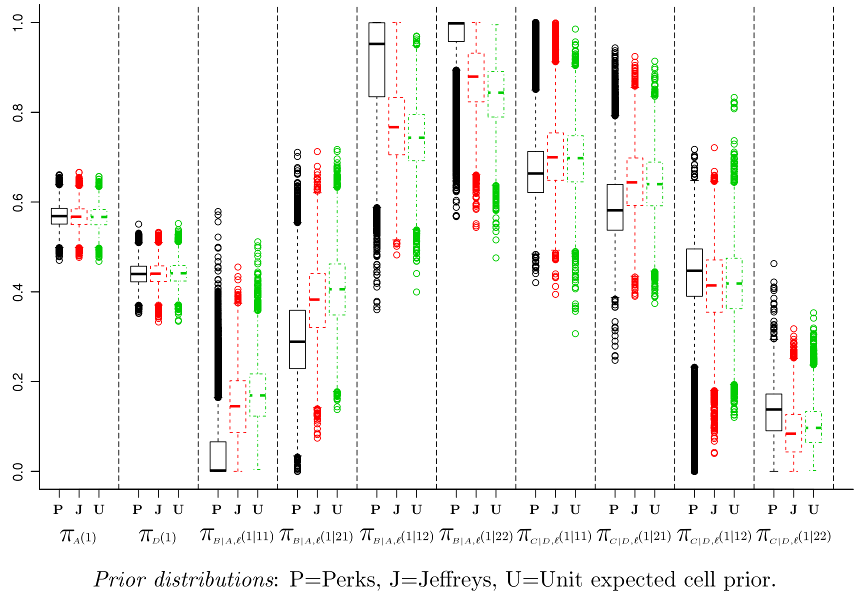


Fig. 2. Boxplots of the posterior distribution of the model parameters $\boldsymbol{\pi}^D$ for the bi-directed 4-chain graph $AB + BC + CD$ fitted on the Coppen's data.

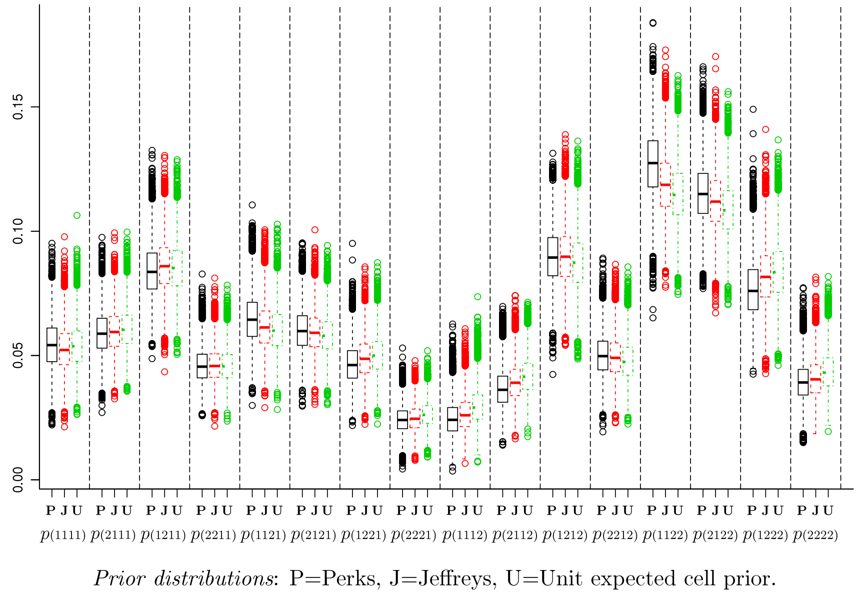


Fig. 3. Boxplots of the posterior distribution of the joint probabilities $\boldsymbol{p}$ for the bi-directed 4-chain graph $AB + BC + CD$ fitted on the Coppen's data.

Illustration on 4-chain $AB + BC + CD$ and chordless 4-cycle $AB + BC + CD + DA$ bi-directed graphs.

Here we present results for the bi-directed 4-chain graph $AB + BC + CD$ and the bi-directed chordless 4-cycle graph $AB + BC + CD + DA$. Both graphs have vertex set $\mathcal{V} = (A, B, C, D)$, the edge set of the first one is $E = \left\{ (\overleftrightarrow{A, B}), (\overleftrightarrow{B, C}), (\overleftrightarrow{C, D}) \right\}$, while for the second we consider the additional edge $(\overleftrightarrow{A, D})$. To obtain posterior summaries, we follow the general approach described in Sections 4.1–4.3, while the marginal likelihood estimator is obtained using the methodology described in Section 4.4. We introduce an additional latent variable ℓ between vertices B and C for the aforementioned bi-directed 4-chain graph and four latent variables, denoted by $\mathcal{L} = \{\ell_1, \ell_2, \ell_3, \ell_4\}$, for the bi-directed chordless 4-cycle graph. By this way, we obtain DAGs which are Markov equivalent over the observed margin to the original bi-directed graphs. Additional details are presented in the Appendix.

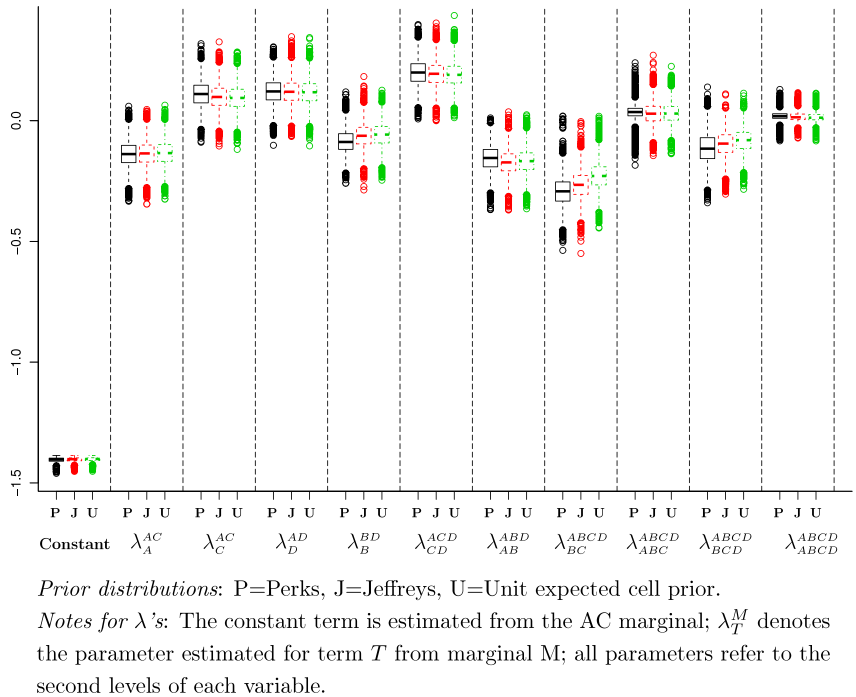


Fig. 4. Boxplots of the posterior distribution of the marginal log-linear parameters $\boldsymbol{\lambda}$ for the bi-directed 4-chain graph $AB + BC + CD$ fitted on the Coppen's data.

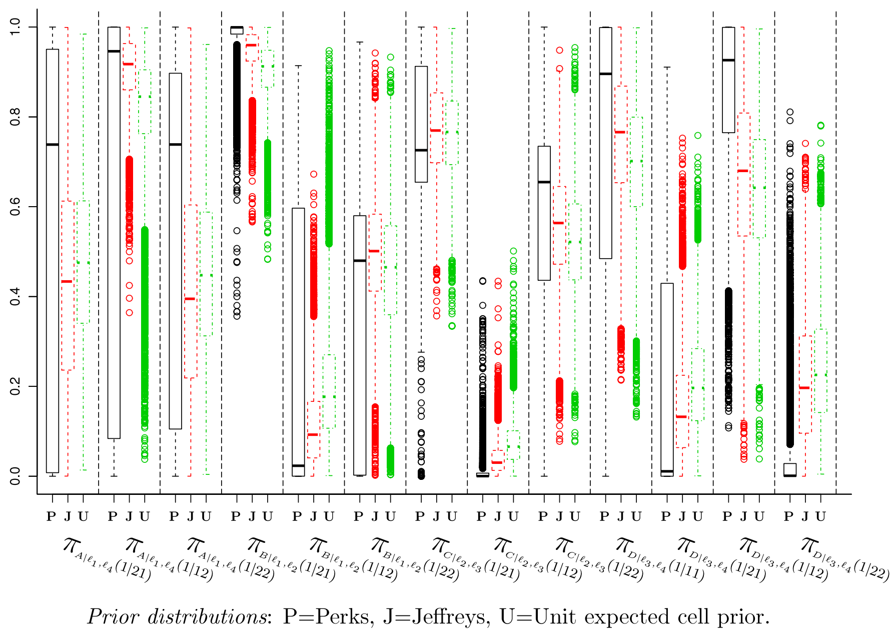


Fig. 5. Boxplots of the posterior distribution of the model parameters $\boldsymbol{\pi}^D$ for the bi-directed chordless 4-cycle graph $AB + BC + CD + DA$ fitted on the Coppen's data.

Figs. 2–4 present box-plots of the posterior distributions for the parameters $\boldsymbol{\pi}^D$, the joint probabilities $\boldsymbol{p}$ of the observed four-way table, and the marginal log-linear parameters $\boldsymbol{\lambda}$ for the fitted model corresponding to the bi-directed 4-chain graph. In these box-plots, although we observe some variability in the posterior distributions of the augmented parameters $\boldsymbol{\pi}^D$, the differences of $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are minor. The corresponding boxplots for the bi-directed chordless 4-cycle are provided in Figs. 5–7. Differences in model parameters $\boldsymbol{\pi}^D$ between the three prior set-ups are more obvious now but still the joint probabilities and the marginal log-linear parameters are close.

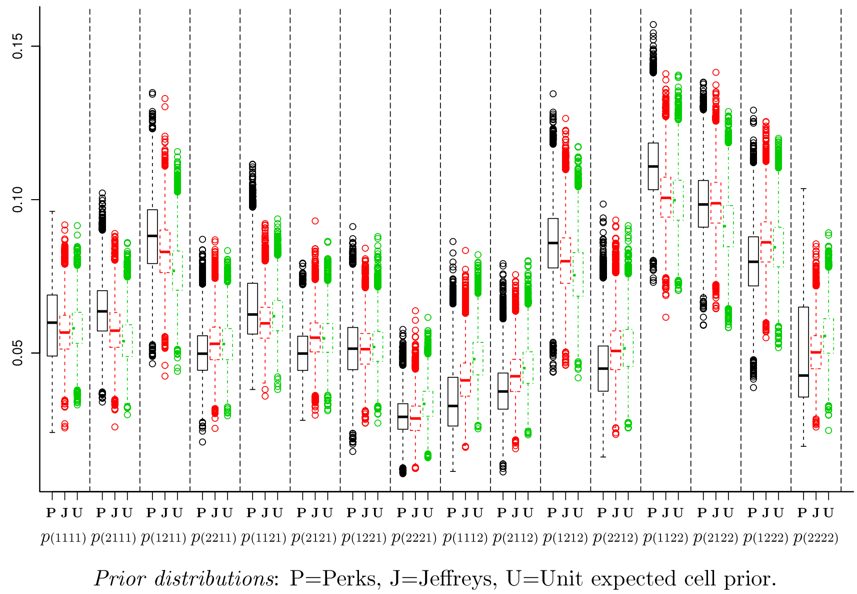


Fig. 6. Boxplots of the posterior distribution of the joint probabilities $\boldsymbol{p}$ for the bi-directed chordless 4-cycle graph $AB + BC + CD + DA$ fitted on the Coppen's data.

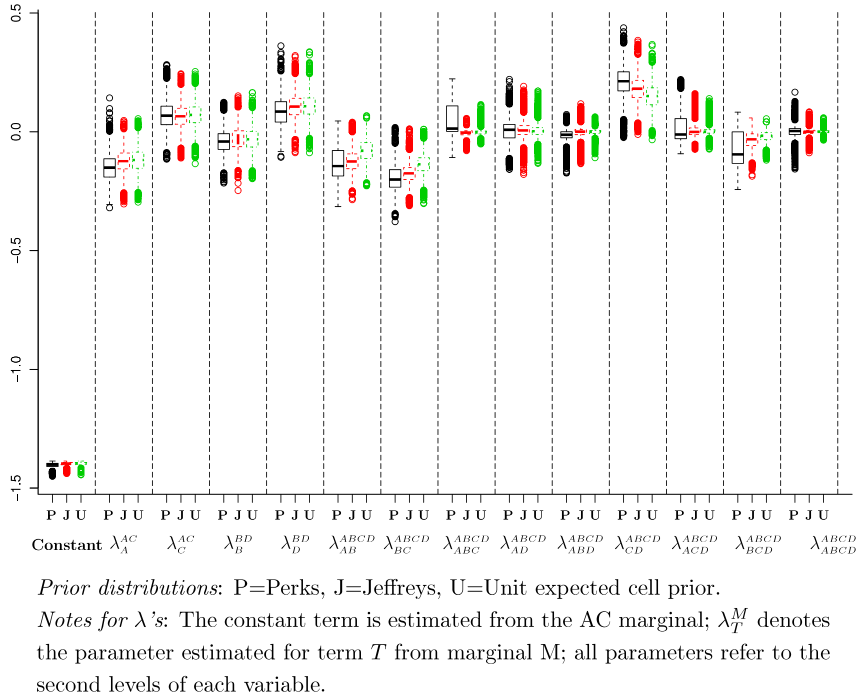


Fig. 7. Boxplots of the posterior distribution of the marginal log-linear parameters $\boldsymbol{\lambda}$ for the bi-directed chordless 4-cycle graph $AB + BC + CD + DA$ fitted on the Coppen's data.

We have implemented a diagnostic test for checking whether the Bell's inequalities are actively involved in the posterior inference of the bi-directed 4-chain graph. According to Ver Steeg and Galstyan (2011, Eq. (8)), if the observed variables are all binary Bell's inequality for the bi-directed 4-chain graph has the following form

$$\sum_{c_i, e_i \in \{0,1\}} P(c_1, c_2 | e_1, e_2) \delta_{(c_1 \bigoplus c_2),(e_1 \times e_2)} \leq 3$$

where $\bigoplus$ is the modulo-2 addition operator and δ is the Kronecker delta. Therefore we have calculated the posterior distributions of the expressions involved in the previous inequality for all prior set-ups using 3000 iterations and additional

Table 2
Marginal log-likelihood estimates for bi-directed 4-chain graph $AB + BC + CD$ fitted on Coppen data under Perks, Jeffreys and unit expected cell priors using different point estimates (averages and standard deviations over 30 samples are reported).

Point estimate π^{*D}	Iterations	Prior set-up		
		Perks $\alpha(i) = 1/2^4$	Jeffreys $\alpha(i) = 1/2$	UEC $\alpha(i) = 1$
Mode	1 000	−64.63 (0.598)	−56.75 (0.225)	−56.68 (0.131)
	10 000	−64.94 (0.535)	−56.68 (0.074)	−56.68 (0.040)
Median	1 000	−64.88 (0.361)	−56.66 (0.156)	−56.69 (0.125)
	10 000	−64.67 (0.098)	−56.7 (0.044)	−56.68 (0.038)
Mean	1 000	−64.94 (0.624)	−56.64 (0.175)	−56.68 (0.119)
	10 000	−64.64 (0.169)	−56.7 (0.046)	−56.68 (0.039)

Table 3
Marginal log-likelihood estimates for bi-directed chordless 4-cycle graph $AB + BC + CD + DA$ fitted on Coppen data under Perks, Jeffreys and unit expected cell priors using different point estimates (averages and standard deviations over 30 samples are reported).

Point estimate π^{*D}	Iterations	Prior set-up		
		Perks $\alpha(i) = 1/2^4$	Jeffreys $\alpha(i) = 1/2$	UEC $\alpha(i) = 1$
Mode	1 000	−69.93 (2.628)	−67.65 (2.460)	−68.08 (1.972)
	10 000	−66.76 (2.056)	−65.79 (2.211)	−66.78 (1.393)
Median	1 000	−61.86 (7.109)	−62.40 (1.999)	−65.77 (2.145)
	10 000	−62.36 (3.244)	−62.85 (1.386)	−66.54 (1.240)
Mean	1 000	−52.59 (8.895)	−61.37 (3.373)	−65.14 (3.723)
	10 000	−55.17 (6.572)	−62.25 (2.220)	−66.66 (1.359)

1000 iterations as burn-in. All posterior distributions are far away from the bounded value of 3. More precisely, all generated values (for all prior set-ups) were not higher than 2.15 indicating that the Bell's inequalities are not actively involved in the posterior inference. Similar results have been obtained for all the other bi-directed 4-chain graphs. The implementation details are provided at the Appendix.

Tables 2 and 3 present the batch mean estimators of the marginal log-likelihood along with the standard deviation of the marginal log-likelihood across 30 batches (i.e. MCMC sub-samples) of 1000 and 10000 iterations. The latter provides an estimate of the Monte Carlo error for the marginal log-likelihood estimate of equivalent size. All results are presented using three point estimates π^{*D} for the model parameters π^D: the mode, the median and the mean.

For the bi-directed 4-chain graph, we observe that for 1000 iteration the Monte-Carlo error is of acceptable size (between 0.12 and 0.63) while for 10 000 iterations the error becomes really low for almost all cases presented in Table 2 (less than 0.1 for all point estimates and prior choices except for the Perks prior using the mode and mean as point estimates with standard deviations 0.53 and 0.17, respectively).

For the bi-directed chordless 4-cycle graph, the Monte Carlo errors are much higher than the corresponding ones in the bi-directed 4-chain graph. This is due to the inclusion of four latent variables, which makes the MCMC slower in terms of convergence. Using the mode as point estimate in the Chib's estimator provides more reliable estimates with Monte Carlo errors (1.97 − 2.63) for 1000 iterations and (1.39 − 2.22) for 10 000 iterations.

Model comparison and evaluation

Table 4 presents results for models with average posterior probability (over 30 MCMC samples) higher than 0.001 under the selected prior distributions. For all models we report the batch mean estimate, its standard error and the standard deviation of the marginal log-likelihood and the corresponding posterior model probabilities over 30 MCMC samples. In all simulations we have used 3000 iterations for the bi-directed 4-chain graphs and 10 000 iterations for chordless 4-cycle graphs. Similar results are presented in Table 5 for the posterior inclusion probabilities of each edge of the graph.

Under all prior set-ups, the maximum a posteriori model (MAP) is the chain $AB + BC + CD$ with probabilities 0.67, 0.91 and 0.42 for Perks, Jeffreys and UEC prior, respectively. For the Perks prior, the relative difference between the MAP model and the second best (bi-directed chordless 4-cycle $AB + BC + CD + DA$) is smaller. The two models cannot be clearly distinguished (in terms of marginal likelihoods or posterior probabilities) when using 3000 or 10 000 iterations for the estimation of the marginal likelihood. This is due to the large Monte Carlo error of the latter model (±2.03 for the log-marginal and ±0.26 for the posterior model probabilities). As a result we cannot safely identify differences between them even if the number of iterations for model $AB + BC + CD + DA$ is increased to 100 000 iterations (marginal log-likelihood Monte Carlo error ≈1.24).

For the rest of the prior distributions, all posterior model probabilities of the best models are accurately estimated. Higher levels of model uncertainty are observed for the UEC prior than the other two set-ups since 15 models have posterior model probability higher than 0.1% for the first in contrast to 5 and 7 models for the other two prior set-ups. For UEC prior, $ABC + CD$ is supported as the second best model with estimated posterior probability 0.17. This model ranked 5th in the Jeffreys prior and 23rd in Perks prior (with probabilities 0.007 and less than 0.001, respectively). Model $A + BC + CD$ is ranked high in

Table 4
Marginal log-likelihood and posterior probabilities (% values) for best models (with estimated posterior probability > 0.001) for the Coppen's data (the batch mean estimates, standard errors, and standard deviations over 30 samples are reported); 3000 and 10 000 iterations were used for the 4-chain and the chordless 4-cycle bi-directed graphs respectively.

Rank	Model	Marginal log-likelihood		Posterior Probability (%)	
		Mean (S.E.)	S.D.	Mean (S.E.)	S.D.
Perks prior $\alpha(i) = 1/2^4$					
1	$AB + BC + CD$ (chain)	−64.75 (0.095)	0.523	67.31 (4.791)	26.24
2	$AB + BC + CD + DA$ (cycle)	−66.21 (0.370)	2.029	30.85 (4.893)	26.80
3	$A + BC + CD$	−68.97 (0.000)	0.000	1.04 (0.113)	0.62
4	$AD + BC + CD$ (chain)	−69.89 (0.083)	0.457	0.49 (0.095)	0.52
5	$A + BCD$	−71.10 (0.000)	0.000	0.12 (0.013)	0.07
Jeffreys prior $\alpha(i) = 1/2$					
1	$AB + BC + CD$ (chain)	−56.74 (0.030)	0.165	91.06 (0.269)	1.47
2	$A + BC + CD$	−59.79 (0.000)	0.000	4.37 (0.126)	0.69
3	$A + BCD$	−60.44 (0.000)	0.000	2.27 (0.065)	0.36
4	$AD + BC + CD$ (chain)	−61.55 (0.040)	0.219	0.77 (0.044)	0.24
5	$ABC + CD$	−61.61 (0.000)	0.000	0.70 (0.020)	0.11
6	$AB + BCD$	−62.64 (0.000)	0.000	0.25 (0.007)	0.04
7	$AB + BC + D$	−62.89 (0.000)	0.000	0.20 (0.006)	0.03
Unit expected cell prior $\alpha(i) = 1$					
1	$AB + BC + CD$ (chain)	−56.68 (0.012)	0.068	42.57 (0.303)	1.66
2	$ABC + CD$	−57.59 (0.000)	0.000	17.14 (0.089)	0.48
3	$A + BC + CD$	−57.81 (0.000)	0.000	13.76 (0.071)	0.39
4	$A + BCD$	−58.11 (0.000)	0.000	10.2 (0.053)	0.29
5	$AB + BCD$	−58.56 (0.000)	0.000	6.55 (0.034)	0.19
6	$ABC + BCD$	−59.24 (0.000)	0.000	3.31 (0.017)	0.09
7	$ABD + BDC$	−59.89 (0.000)	0.000	1.72 (0.009)	0.05
8	$ACB + ACD$	−60.5 (0.000)	0.000	0.94 (0.005)	0.03
9	$AC + BC + CD$	−60.77 (0.000)	0.000	0.71 (0.004)	0.02
10	$ABCD$	−60.8 (0.000)	0.000	0.69 (0.004)	0.02
11	$AD + BC + CD$ (chain)	−60.91 (0.018)	0.099	0.63 (0.013)	0.07
12	$AB + BC + D$	−60.95 (0.000)	0.000	0.60 (0.003)	0.02
13	$AC + BCD$	−61.07 (0.000)	0.000	0.53 (0.003)	0.01
14	$ABC + D$	−61.86 (0.000)	0.000	0.24 (0.001)	0.01
15	$ACD + BC$	−62.34 (0.000)	0.000	0.15 (0.001)	0.00

Table 5
Posterior inclusion probabilities (% values) for each edge of the bi-directed 4-way graph for the Coppen's data (the batch mean estimates (standard errors) over 30 samples are reported; 3000 and 10 000 iterations were used for the 4-chain and the chordless 4-cycle bi-directed graphs respectively).

Edge	Prior set-up		
	Perks $\alpha(i) = 1/16$	Jeffreys $\alpha(i) = 1/2$	UEC $\alpha(i) = 1$
AB	98.2 (0.22)	92.5 (0.22)	73.8 (0.14)
AC	0.0 (0.00)	0.9 (0.03)	23.8 (0.12)
AD	31.4 (4.85)	0.9 (0.06)	4.3 (0.03)
BC	100.0 (0.00)	100.0 (0.00)	100.0 (0.00)
BD	0.1 (0.01)	2.7 (0.08)	23.1 (0.12)
CD	99.8 (0.02)	99.7 (0.01)	99.0 (0.01)

all prior set-ups: 3rd, 2nd and 3rd for Perks, Jeffreys and UEC priors, respectively with posterior probabilities 0.010, 0.044 and 0.138.

Finally, Table 5 presents the summary statistics of the inclusion probabilities of each edge giving a clearer picture of the edges representing important dependences. According to this table, edges AB, BC and CD should be included in the finally selected graph with posterior inclusion probabilities at least 0.73, 1.00 and 0.99, respectively. Edge AD is mildly supported only under the Perks prior (inclusion probability 0.31 ± 0.05) while edges AC and BD are only weakly supported with inclusion probabilities around 0.23 for the UEC prior. Hence, the bi-directed 4-chain graph $AB + BC + CD$ is indicated as the median probability model (Barbieri and Berger, 2004) in all three prior set-ups.

6. Discussion and final comments

In this article, we have presented a novel Bayesian approach for the analysis of discrete graphical models of marginal independence. We have exploited the connection between bi-directed graphs and DAGs by expressing them as models of conditional association. In this way, it was feasible to apply the recursive factorisation of the joint probability distribution of DAGs, and use suitable conjugate prior distributions. Posterior distributions were either readily available for parameters of bi-directed graphs with direct DAG representation, or estimated using a Gibbs sampler obtained by a data augmentation scheme. Chib's estimator was used to calculate the marginal likelihood of models without a direct DAG representation. Moreover, specific details were provided for the 4-way case along with an illustration in a well known dataset.

It worth noting that, the representation in term of augmented DAG is not unique for all graphs. In this work, we have considered only one of the possible equivalent DAGs. Nevertheless, since the prior distributions are compatible across models, the posterior distributions and the marginal likelihoods will not depend on this choice; see Buntine (1991) and Heckerman et al. (1995).

Even though the methodology presented here is general and can be applied for models of any dimension, its applicability to high dimensional contingency tables may be problematic in practice. This is due to the elevated number of latent variables that should be included in the augmented DAG. Therefore, for high dimensional problems, a more efficient methodology may be required. An alternative approach to estimate posterior model probabilities, on which we are working on, is to consider an appropriate trans-dimensional MCMC algorithm, see Sisson (2005) with emphasis given in reversible jump MCMC; Green (1995).

Finally, another interesting direction that we are currently considering, is to work directly with the marginal log-linear parameterisation $\boldsymbol{\lambda}$ defined by (8). In this case, a conjugate analysis is not feasible, and a more complicated approach is necessary. In this direction, two alternative approaches are under investigation: (a) a MCMC based directly on simulating the parameters of each marginal association log-linear parameters following the approach proposed by Knuiman and Speed (1988) and Dellaportas and Forster (1999), and (b) a Metropolis–Hastings algorithm with proposals based on the probability parameterisation we have considered in this article. A possible disadvantage of the first approach is that in each iteration of the MCMC sampler we need to implement iterative methods to calculate the cell probabilities and thus the calculation of the model likelihood will reduce the efficiency of the algorithm. Finally, implementing RJMCMC algorithm for the selection of the graphical structure seems a natural conclusion of this approach.

Acknowledgements

We would like to thank the two anonymous reviewers for their helpful suggestions, and Monia Lupparelli for fruitful discussions about graphical models of marginal independence. This work was partially supported by MIUR, Rome, under project PRIN 2005132307, University of Pavia, and by the "Basic research funding program" of the Research Centre of the Athens University of Economics and Business.

Appendix A. Posterior inference for a bi-directed 4-chain graph

Here we provide specific details for the implementation of Bayesian inference for a bi-directed 4-chain graph. We consider the graph with vertex set $\mathcal{V} = \{e_1, c_1, c_2, e_2\}$ and edge set $E = \left\{(\overleftrightarrow{e_1, c_1}), (\overleftrightarrow{c_1, c_2}), (\overleftrightarrow{c_2, e_2})\right\}$ represented in Fig. A.8(a). This graph is Markov equivalent over the observed margin to a DAG with an additional latent variable ℓ added between c_1 and c_2, see Fig. A.8(b). The joint probabilities (needed in the likelihood) for the original 4-way table are given by

$$p(i) = \sum_{i_\ell \in \mathcal{I}_\mathcal{L}} p^{\mathcal{A}}(i, i_\ell) = \sum_{i_\ell \in \mathcal{I}_\mathcal{L}} \left\{ \pi_\ell(i_\ell) \prod_{k=1}^{2} \pi_{e_k}(i_{e_k})\, \pi_{c_k|e_k,\ell}(i_{c_k}|i_{e_k}, i_\ell) \right\} \tag{A.1}$$

where $\mathcal{A} = \mathcal{V} \cup \mathcal{L} = \{e_1, c_1, c_2, e_2, \ell\}$ and $\mathcal{L} = \{\ell\}$.

The number of parameters in the above augmented model is equal to

$$\mathrm{P}^{\mathrm{D}} = \sum_{k=1}^{2} \left(|\mathcal{I}_{e_k}| - 1\right) + \sum_{k=1}^{2} \left(|\mathcal{I}_{c_k}| - 1\right) |\mathcal{I}_{e_k}|\, |\mathcal{I}_\ell| + (|\mathcal{I}_\ell| - 1)$$

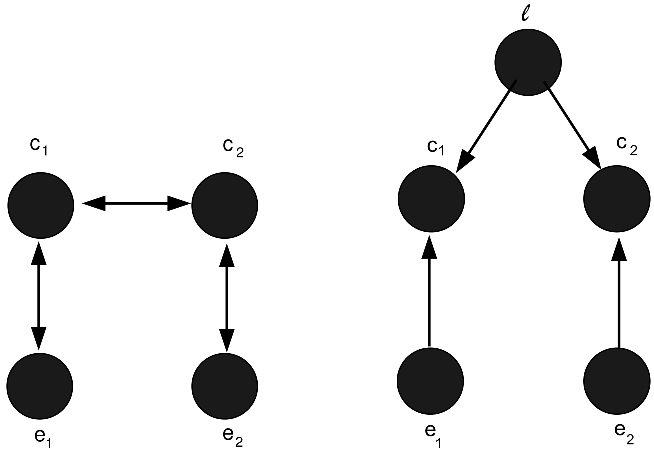
while the original model has

$$\mathrm{P}^{\mathrm{G}} = \prod_{k=1}^{2} \left(|\mathcal{I}_{e_k}|\, |\mathcal{I}_{c_k}| + 1\right) - \left(\prod_{k=1}^{2} |\mathcal{I}_{e_k}|\right) \left(\sum_{k=1}^{2} |\mathcal{I}_{c_k}| - 1\right) - 3$$

parameters. The constraints are set using the approach described in Section 4.2. For the 2^4 example implemented in Section 5, we have $\mathrm{P}^{\mathrm{G}} = 10$ and $\mathrm{P}^{\mathrm{D}} = 11$ parameters when $|\mathcal{I}_\ell| = 2$. Thus, only one constraint is needed. Here we have considered $\pi_\ell(i_\ell = 1) = 1/2$.

A.1. Gibbs sampling for a bi-directed 4-chain graph

The Gibbs sampling described in Section 4.1 is implemented as follows:



(a) Bi-directed graph. (b) DAG representation.

Fig. A.8. Bi-directed 4-chain graph and the corresponding Markov equivalent DAG over the observed margin.

1. Generate $n^{\mathcal{A}}(i, i_\ell) \sim Multinomial(\widetilde{\boldsymbol{p}}(i), n)$ with $\widetilde{\boldsymbol{p}}(i)$ being a vector of length $|\mathcal{I}_\ell|$ and elements
$$\widetilde{\boldsymbol{p}}(i, i_\ell) = \frac{p^{\mathcal{A}}(i, i_\ell)}{\sum\limits_{i'_\ell \in \mathcal{I}_{\mathcal{L}}} p^{\mathcal{A}}(i, i'_\ell)}$$
for $i_\ell \in \mathcal{I}_\ell$ and any $i \in \mathcal{I}$.
2. Generate $\pi_{e_k} \sim Dirichlet\left(\boldsymbol{n}_{e_k} + \boldsymbol{\alpha}_{e_k}\right)$ for $k = 1, 2$.
3. Generate $\pi_\ell \sim Dirichlet\left(\boldsymbol{n}_\ell^{\mathcal{A}} + \boldsymbol{\alpha}_\ell\right)$.
4. For $k = 1, 2$, $i_{e_k} \in \mathcal{I}_{e_k}$ and $i_\ell \in \mathcal{I}_\ell$, generate $\boldsymbol{\pi}_{c_k|i_{e_k},i_\ell} \sim Dirichlet\left(\boldsymbol{n}^{\mathcal{A}}_{c_k|i_{e_k},i_\ell} + \boldsymbol{\alpha}_{c_k|i_{e_k},i_\ell}\right)$.

The above MCMC implements the model with no constraints on $\boldsymbol{\pi}^D$. The constrained version of the model can be estimated in a similar way but in steps 3 and 4 the corresponding conditional Dirichlet distributions must be used instead. For the binary case presented in Section 5, step 3 should be skipped since $\pi_\ell(i_\ell) = 1/2$.

A.2. Marginal likelihood computation for a bi-directed 4-chain graph

For the estimation of the marginal likelihood, we use the Chib (1995) estimator as described in Section 4.4 using the output of the MCMC described in Appendix A.1 for the constrained version of the model. As $\boldsymbol{\pi}^{*D}$ we use three different points: the posterior mode, the posterior median and the posterior mean. The posterior mode is approximated via the MCMC output. Although using the MCMC is not the most efficient way to estimate the posterior mode, here the loss of the precision is not essential since the Chib's marginal likelihood estimator works well for any point of high posterior density.

The prior is simply the product of independent Dirichlet probability densities for each unconstrained component of $\boldsymbol{\pi}^D$ evaluated at $\boldsymbol{\pi}^{*D}$ The posterior ordinate $f(\boldsymbol{\pi}^{*D}|\boldsymbol{y})$ is estimated from the Gibbs sampling output using the estimator

$$\widehat{f}\left(\boldsymbol{\pi}^{*D}|\boldsymbol{y}\right) = \prod_{k=1}^{2} f_{Di}\left(\boldsymbol{\pi}^*_{e_k}; \boldsymbol{n}_{e_k} + \boldsymbol{\alpha}_{e_k}\right) \times \frac{1}{T}\sum_{t=1}^{T}\left\{ f_{Di}\left(\boldsymbol{\pi}^*_\ell; \boldsymbol{n}_\ell^{\mathcal{A}(t)} + \boldsymbol{\alpha}_\ell\right) \prod_{k=1}^{2}\prod_{i_{e_k}\in\mathcal{I}_{e_k}}\prod_{i_\ell\in\mathcal{I}_{v_\ell}} f_{Di}\left(\boldsymbol{\pi}^*_{c_k|e_k,\ell}; \boldsymbol{n}^{\mathcal{A}(t)}_{c_k|\,e_k,\ell} + \boldsymbol{\alpha}_{c_k|\,e_k,\ell}\right)\right\},$$

where $\boldsymbol{n}^{\mathcal{A}(t)}_{c_k|i_{e_k},i_\ell}$ is a vector of frequency data with elements $n^{\mathcal{A}(t)}_{c_k,e_k,\ell}(i_{c_k}, i_{e_k}, i_\ell)$ for $i_{c_k} \in \mathcal{I}_{c_k}$ and given i_{e_k}, i_ℓ. The superscript (t) refers to the t-th iteration and $n^{\mathcal{A}(t)}(i)$ for $i \in \mathcal{I}_{\mathcal{A}}$ refers to the augmented 5-way table after the introduction of the latent factor ℓ generated at the t-th iteration. Note that, in the equation above the densities must be replaced by the corresponding conditional Dirichlet densities if some parameters are constrained. For the binary case considered in Section 5 with $p(i_\ell = 1) = 1/2$, $f_{Di}\left(\boldsymbol{\pi}^*_\ell; \boldsymbol{n}_\ell^{\mathcal{A}(t)} + \boldsymbol{\alpha}_\ell\right)$ must be eliminated from this expression. Finally, following Neal (1998) and Marin and Robert (2008), the marginal log-likelihood must be corrected by adding a factor equal to $\log(|\mathcal{I}_\ell|!)$ which results to $\log 2 = 0.693$ for the case of a binary latent variable.

A.3. Calculations for checking Bell's inequality for a bi-directed 4-chain graph

If the observed variables are all binary the Bell's inequality for a bi-directed 4-chain graph has the following form

$$\begin{aligned}&P(c_1=0,c_2=0|e_1=0,e_2=0)+P(c_1=1,c_2=1|e_1=0,e_2=0)\\&\quad+P(c_1=0,c_2=0|e_1=1,e_2=0)+P(c_1=1,c_2=1|e_1=1,e_2=0)\\&\quad+P(c_1=0,c_2=0|e_1=0,e_2=1)+P(c_1=1,c_2=1|e_1=0,e_2=1)\\&\quad+P(c_1=1,c_2=0|e_1=1,e_2=1)+P(c_1=0,c_2=1|e_1=1,e_2=1)\leq 3.\end{aligned} \tag{A.2}$$

Assuming that from an MCMC, a sample of $\left(\boldsymbol{\pi}^{(t)}_{e_1},\boldsymbol{\pi}^{(t)}_{e_2},\boldsymbol{\pi}^{(t)}_{\ell},\boldsymbol{\pi}^{(t)}_{c_1|e_1,\ell},\boldsymbol{\pi}^{(t)}_{c_2|e_2,\ell}\right)$ for $t=1,\ldots,T$ is available, then, for each iteration t, we perform the following computations

1. Calculate the joint probabilities $\boldsymbol{p}^{(t)}$ for the observed table for each iteration using Eq. (A.1).
2. Calculate

$$\pi^{(t)}_{\mathbb{C}|\mathbb{E}}(i_{\mathbb{C}}|i_{\mathbb{E}})=\frac{p^{(t)}(i_{\mathbb{C}},i_{\mathbb{E}})}{\sum_{i\in\mathcal{I}_{\mathbb{C}}}p^{(t)}(i_{\mathbb{C}},i_{\mathbb{E}})}$$

   for the levels $i_{\mathbb{C}}$ and $i_{\mathbb{E}}$ involved in (A.2); where $\mathbb{C}=\{c_1,c_2\}$ and $\mathbb{E}=\{e_1,e_2\}$ are the sets of corner and ending nodes respectively.
3. Calculate the expression $B^{(t)}$ involved in Bell's inequality (A.2) for each iteration t by

$$\begin{aligned}B^{(t)}=&\sum_{i_{\mathbb{E}}\neq(1,1)}\pi^{(t)}_{\mathbb{C}|\mathbb{E}}(i_{\mathbb{C}}=\mathbf{0}|i_{\mathbb{E}})+\sum_{i_{\mathbb{E}}\neq(1,1)}\pi^{(t)}_{\mathbb{C}|\mathbb{E}}(i_{\mathbb{C}}=\mathbf{1},i_{c_2}=1|i_{\mathbb{E}})\\&+\pi^{(t)}_{\mathbb{C}|\mathbb{E}}(i_{c_1}=1,i_{c_2}=0|i_{\mathbb{E}}=\mathbf{1})+\pi^{(t)}_{\mathbb{C}|\mathbb{E}}(i_{c_1}=0,i_{c_2}=1|i_{\mathbb{E}}=\mathbf{1});\end{aligned}$$

   where $\mathbf{0}=(0,0)^T\mathbf{1}=(1,1)$.

Then we can use $\left(B^{(t)},t=1,\ldots,T\right)$ to assess the posterior distribution of the expression involved in Bell's inequality. Evidence that the inequality is active, affecting the posterior inference will be suggested only if the posterior distribution and the corresponding mode is close to the upper bound of 3.

Appendix B. Posterior inference for a bi-directed chordless 4-cycle graph

Here we provide the details for the implementation of the Bayesian inference for bi-directed chordless 4-cycle graphs with vertex set $\mathcal{V}=\{c_1,c_2,c_3,c_4\}$ and edge set $E=\left\{(\overleftrightarrow{c_1,c_2}),(\overleftrightarrow{c_2,c_3}),(\overleftrightarrow{c_3,c_4}),(\overleftrightarrow{c_4,c_1})\right\}$ represented in Fig. B.9(a). This graph is Markov equivalent over the observed margin to a DAG with four additional latent variables $\mathcal{L}=\{\ell_1,\ell_2,\ell_3,\ell_4\}$ (see Fig. B.9(b)) with parameters π_{ℓ_k} and $\pi_{c_k|\ell_{k-1},\ell_k}$ for $k=1,2,3,4$ and $\ell_0=\ell_4$. The joint probabilities for the original 4-way table are given by

$$p(i)=\sum_{i_L\in\mathcal{I}_{\mathcal{L}}}p^{\mathcal{A}}(i,i_{\mathcal{L}})=\sum_{i_L\in\mathcal{I}_{\mathcal{L}}}\prod_{k=1}^{4}\pi_{\ell_k}(i_{\ell_k})\pi_{c_k|\ell_{k-1},\ell_k}(i_{c_k}|i_{\ell_{k-1}},i_{\ell_k}) \tag{B.1}$$

where $\mathcal{A}=\mathcal{V}\cup\mathcal{L}=\{c_1,c_2,c_3,e_4,\ell_1,\ell_2,\ell_3,\ell_4\}$ and $i_L=\{i_{\ell_1},i_{\ell_2},i_{\ell_3},i_{\ell_4}\}$.

B.1. Gibbs sampling for a bi-directed chordless 4-cycle graph

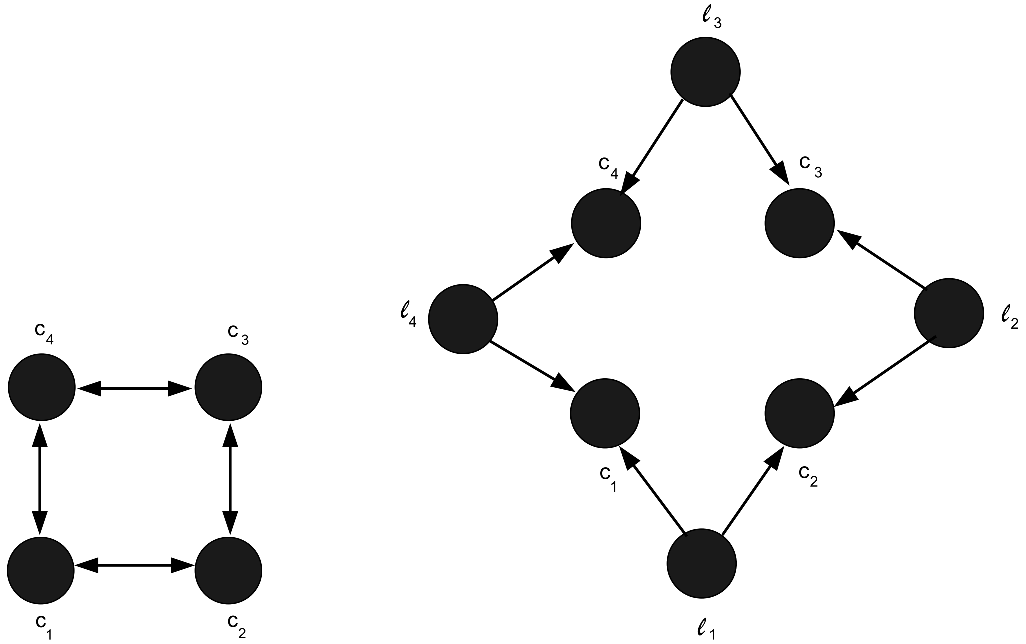
The MCMC scheme is similar to the one presented for the bi-directed 4-chain model and it can be summarised by the following steps:

1. Generate $n^{\mathcal{A}}(i,i_\ell)\sim Multinomial\,(\widetilde{\boldsymbol{p}}(i),n)$ with $\widetilde{\boldsymbol{p}}(i)$ given by (7).
2. For $c_k\in\mathcal{V}$, $i_{e_k}\in\mathcal{I}_{e_k}$, $i_\ell\in\mathcal{I}_\ell$, generate

$$\boldsymbol{\pi}_{c_k|i_{\ell_{k-1}},i_{\ell_k}}\sim Dirichlet\left(\boldsymbol{n}^{\mathcal{A}}_{c_k|i_{\ell_{k-1}},i_{\ell_k}}+\boldsymbol{\alpha}_{c_k|i_{\ell_{k-1}},i_{\ell_k}}\right)$$

   with $\ell_0=\ell_4$.
3. For $\ell_k\in\mathcal{L}$, generate $\pi_{\ell_k}\sim Dirichlet\left(\boldsymbol{n}^{\mathcal{A}}_{\ell_k}+\boldsymbol{\alpha}_{\ell_k}\right)$.

Similarly to Appendix A.1, the above sampler is for the unconstrained model. If the constrained version of the model is considered then steps 2 and 3 must be changed accordingly to accommodate these constraints.



(a) Bi-directed graph. (b) DAG representation.

Fig. B.9. Bi-directed chordless 4-cycle graph and the corresponding Markov equivalent DAG over the observed margin.

The number of parameters in the bi-directed chordless 4-cycle graph presented in Fig. B.9(a) is given by

$$\mathrm{P}^{\mathrm{G}} = \prod_{k=1}^{4} |\mathcal{I}_{c_k}| - \sum_{k=1}^{2} \left\{ \prod_{j=0}^{1} \left(|\mathcal{I}_{c_{k+2j}}| - 1 \right) \right\} - 1$$

while for the corresponding DAG, the number of parameters is given by

$$\mathrm{P}^{\mathrm{D}} = \sum_{k=1}^{4} \left(|\mathcal{I}_{\ell_k}| - 1 \right) + \sum_{k=1}^{4} \left(|\mathcal{I}_{c_k}| - 1 \right) |\mathcal{I}_{\ell_{k-1}}| \, |\mathcal{I}_{\ell_k}|.$$

For bi-directed chordless 4-cycle graphs with binary variables, we need to impose seven constraints since $\mathrm{P}^{\mathrm{G}} = 13$ and $\mathrm{P}^{\mathrm{D}} = 20$ for binary latent variables. In the illustration of Section 5, we use $\pi_{\ell_k}(i_{\ell_k}) = 1/2$ for all $i_{\ell_k} \in \mathcal{I}_{\ell_k}$ and $k = 1, 2, 3, 4$. Moreover, we set $\pi_{c_k|\ell_{k-1},\ell_k}(i_{c_k}|i_{\ell_{k-1}}, i_{\ell_k}) = 1/2$ for $k = 1, 2, 3$.

B.2. Marginal likelihood computation for chordless bi-directed 4-cycle graphs

The prior and the posterior ordinate involved in the estimation of the marginal likelihood using the estimator of Chib (1995) are now given by

$$\log f\left(\boldsymbol{\pi}^{*D}\right) = \sum_{k=1}^{4} \left\{ \log f_{Di}\left(\boldsymbol{\pi}^*_{\ell_k}; \boldsymbol{\alpha}_{\ell_k}\right) + \sum_{i_{\ell_{k-1}} \in \mathcal{I}_{\ell_{k-1}}} \sum_{i_{\ell_k} \in \mathcal{I}_{\ell_k}} \log f_{Di}\left(\boldsymbol{\pi}^*_{c_k|i_{\ell_{k-1}},i_{\ell_k}}; \boldsymbol{\alpha}_{c_k|i_{\ell_{k-1}},i_{\ell_k}}\right) \right\}$$

and

$$\widehat{f}\left(\boldsymbol{\pi}^{*D}|\boldsymbol{y}\right) = \frac{1}{T} \sum_{t=1}^{T} \left\{ \prod_{k=1}^{4} f_{Di}\left(\boldsymbol{\pi}^*_{\ell_k}; \boldsymbol{\alpha}_{\ell_k}\right) \times \prod_{k=1}^{4} \prod_{i_{\ell_{k-1}} \in \mathcal{I}_{\ell_{k-1}}} \prod_{i_{\ell_k} \in \mathcal{I}_{v_{\ell_k}}} f_{Di}\left(\boldsymbol{\pi}^*_{c_k|i_{\ell_{k-1}},i_{\ell_k}}; \boldsymbol{n}^{\mathcal{A}(t)}_{c_k|i_{\ell_{k-1}},i_{\ell_k}} + \boldsymbol{\alpha}_{c_k|i_{\ell_{k-1}},i_{\ell_k}}\right) \right\},$$

respectively. Similarly to the implementation of the method for the 4-chain graph, in the expression above the densities must be substituted by the corresponding induced conditional Dirichlet densities if some parameters are constrained. Finally, to account for the label switching problem, the correction term that must be added in the marginal likelihood is equal to $\sum_{k=1}^{4} \log(|\mathcal{I}_{\ell_k}|!)$ which results to $4 \log 2 = 2.77$ for the case of binary latent variables.

References

Barbieri, M.M., Berger, J., 2004. Optimal predictive model selection. Annals of Statistics 32, 870–897.
Bartlett, M.S., 1957. A comment on Lindley's statistical paradox. Biometrika 44, 533–534.
Bartolucci, F., Scaccia, L., Farcomeni, A., 2012. Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data. Computational Statistics and Data Analysis 56, 4067–4080.
Bergsma, W.P., 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., Bonissone, P. (Eds.), Proceedings of the Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 52–60.
Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.
Consonni, G., Veronese, P., 2008. Compatibility of prior specifications across linear models. Statistical Science 23, 332–353.
Coppen, A., 1966. The Mark–Nyman temperament scale: an English translation. British Journal of Medical Psychology 39, 55–59.
Cox, D.R., Wermuth, N., 1993. Linear dependencies represented by chain graphs. Statistical Science 8 (with discussion) 204–218, 247–277.
Dawid, A.P., Lauritzen, S.L., 2000. Compatible prior distributions. In: George, E.I. (Ed.), Bayesian methods with applications to science policy and official statistics. The Sixth World Meeting of the International Society for Bayesian Analysis. pp. 109–118. http://www.stat.cmu.edu/ISBA/index.html.
Dellaportas, P., Forster, J.J., 1999. Markov Chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86, 615–633.
Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society B 56, 363–375.
Drton, M., Richardson, T.S., 2008a. Binary models for marginal independence. Journal of the Royal Statistical Society B 70, 287–309.
Drton, M., Richardson, T.S., 2008b. Graphical methods for efficient likelihood inference in Gaussian covariance models. Journal of Machine Learning Research 9, 893–914.
Evans, R.J., 2012. Graphical mathods for ineuqality constraints in marginalised DAGs. IEEE International Workshop on Machine Learning for Signal Processing 1–6.
Evans, R.J., Richardson, T.S., 2013. Marginal log-linear parameters for graphical Markov models. Journal of the Royal Statistical Society B (in press).
Frigyik, B.A., Kapila, A., Gupta, M.R., 2010. Introduction to the Dirichlet distribution and related processes. Technical Report. UWEETR-2010-0006. Department of Electrical Engineering. University of Washington.
Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal American Statistical Association 96, 194–209.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Heckerman, D., Geiger, D., Chickering, D., 1995. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20, 194–243.
Jasra, A., Holmes, C.C., Stephens, D.A., 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20, 50–67.
Kang, C., Tian, J., 2012. Inequality constraints in causal models with hidden variables. arXiv:1206.6829.
Khare, K., Rajaratnam, B., 2011. Wishart distributions for decomposable covariance graph models. Annals of Statistics 39, 514–555.
Knuiman, M.W., Speed, T.P., 1988. Incorporating prior information into the analysis of contingency tables. Biometrics 44, 1061–1071.
Letac, G., Massam, H., 2007. Wishart distributions for decomposable graphs. The Annals of Statistics 35, 1278–1323.
Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.
Lupparelli, M., 2006. Graphical models of marginal independence for categorical variables. Ph.D. Thesis. University of Florence.
Lupparelli, M., Marchetti, G.M., Bergsma, W.P., 2009. Parameterization and fitting of discrete bi-directed graph models. Scandinavian Journal of Statistics 36, 559–576.
Marin, J.-M., Robert, C.P., 2008. Approximating the marginal likelihood in mixture models. Bulletin of the Indian Chapter of ISBA V (1), 2–7. Also available as arXiv:0804.2414 at http://arxiv.org/abs/0804.2414.
Neal, R.M., 1998. Erroneous results in marginal likelihood from the Gibbs output (with associated comments written 16 March 1999). Technical Report. Department of Statistics and Department of Computer Science. University of Toronto. Canada. Available at http://www.cs.toronto.edu/~radford/chib-letter.html.
Ntzoufras, I., Tarantola, C., 2012. Bayesian analysis of graphical models of marginal independence for three way contingency tables. Technical Report. Department of Political Economy and Quantitative Methods. University of Pavia.
Papastamoulis, P., Iliopoulos, G., 2010. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. Journal of Computational and Graphical Statistics 19, 313–331.
Pearl, J., Wermuth, N., 1994. When can association graphs admit a causal interpretation? In: Cheesman, P., Oldford, W. (Eds.), Models and Data, Artificial Intelligence and Statistics IV. Springer, New York, pp. 205–214.
Perks, W., 1947. Some observations on inverse probability including a new indifference rule. Journal of the Institute of Actuaries 73, 285–334.
Richardson, T.S., 2003. Markov property for acyclic directed mixed graphs. Scandinavian Journal of Statistics 30, 145–157.
Roverato, A., Consonni, G., 2004. Compatible prior distributions for DAG models. Journal of the Royal Statistical Society B 66, 47–61.
Roverato, A., Lupparelli, M., La Rocca, L., 2013. Log-mean linear models for binary data. Biometrika http://dx.doi.org/10.1093/biomet/ass080.
Rudas, T., Bergsma, W.P., 2004. On applications of marginal models for categorical data. Metron LXII, 1–25.
Silva, R., Ghahramani, Z., 2009a. The hidden life of latent variables: Bayesian learning with mixed graph models. Journal of Machine Learning Research 10, 1187–1238.
Silva, R., Ghahramani, Z., 2009b. Factorial mixture of Gaussians and the marginal independence model. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS). Clearwater Beach, Florida, USA.
Sisson, S.A., 2005. Trans-dimensional Markov chains: a decade of progress and future perspectives. Journal of the Amarican Statistical association 100, 1077–1089.
Stephens, M., 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society B 62, 795–809.
Tian, J., Pearl, J., 2002. On the testable implications of causal models with hidden variables. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 519–552.
Ver Steeg, G., Galstyan, A., 2011. A sequence of relaxation constraining hidden variable models. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. AUAI Press, pp. 717–726.
Wermuth, N., 1976. Model search among multiplicative models. Biometrics 32, 253–263.
Yao, W., 2012. Model based labeling for mixture models. Statistics and Computing 22, 337–347.