Probability Based Independence Sampler for Bayesian Quantitative Learning in Graphical Log-Linear Marginal Models

Ioannis Ntzoufras∗
Department of Statistics, Athens University of Economics and Business, Greece

Claudia Tarantola
Department of Economics and Management, University of Pavia, Italy

Monia Lupparelli
Department of Statistical Sciences, University of Bologna, Italy

July 1, 2016

Abstract

Bayesian methods for graphical log-linear marginal models has not been developed in the same extend as traditional frequentist approaches. In this work, we introduce a novel Bayesian approach for quantitative learning for such models. Posterior distributions cannot be obtained analytically and MCMC methods are required. Well-defined models of marginal independence require parameter values that lead to compatible marginal probabilities. This is ensured if we move within the space of variation independent parameterisations which makes arduous the implementation of any MCMC scheme. Furthermore, the likelihood of the model cannot be analytically expressed as a function of the marginal log-linear interactions. We construct a novel, fully automatic and efficient MCMC strategy for quantitative learning for graphical log-linear marginal models that handles these problems. While the prior is expressed in terms of the marginal log-linear parameters, we build an MCMC algorithm which employs a proposal on the probability parameter space. By this strategy we achieve to move within the target space of variation independent parameterisations. Moreover, we can exploit the conditional conjugate setups build on probability parameters to construct efficient proposal distributions. The proposed methodology is illustrated using a popular four-way dataset.

∗Address for correspondence: Ioannis Ntzoufras, Department of Statistics, Athens University of Economics and Business, Athens, Greece. E-mail: ntzoufras@aueb.gr
Keywords: Graphical models, Marginal Log-Linear Parameterisation, Markov Chain Monte Carlo Computation.

1 Introduction

Statistical models which impose restrictions on marginal distributions of categorical data have received considerable attention especially in social and economic sciences; see, for example, in Bergsma et al. (2009). A particular appealing class is that of log-linear marginal models introduced by Bergsma and Rudas (2002), that includes as special cases log-linear and multivariate logistic models. The marginal log-linear interactions are estimated using the frequencies of appropriate marginal contingency table, and expressed in terms of log-odds ratios. This setup is important in cases where information is available for specific marginal associations via odds ratios (i.e. marginal log-linear interactions) or when partial information (i.e. marginals) is available.

Log-linear marginal models have been used to provide parameterisations for discrete graphical models; see Lupparelli et al. (2009), Rudas et al. (2010) and Evans and Richardson (2013). In particular, Lupparelli et al. (2009) used them to define a parameterisation for discrete graphical models of marginal independence represented by a bi-directed graph. The absence of an edge in the bi-directed graph indicates marginal independence, and the corresponding marginal log-linear interactions (i.e. the corresponding log-odds ratio) are constrained to zero.

Despite the increasing interest in the literature for graphical log-linear marginal models, Bayesian analysis has not been developed as much as traditional methods. Some context specific results have been presented by e.g. Silva and Ghahramani (2009), Bartolucci et al. (2012) and Ntzoufras and Tarantola (2013).

In order to have a well-defined model of marginal independence, we need to obtain parameter values that lead to compatible marginal probabilities. This is ensured if we move within the space of variation independent parameterisations. Ignoring restrictions imposed by variation independence may cause severe problems leading to misleading parameter values and false interpretation of marginal dependencies (Rudas and Bergsma, 2004).

In terms of Bayesian inference, the aim is to calculate the posterior distribution within the restricted space of variation independence parameterisations. Since this is a cumbersome distribution, no conjugate analysis is available and the therefore Markov chain Monte Carlo (MCMC) methods must be employed. The implementation of any MCMC within the space of variation independence parameterisations is not a straightforward task. Furthermore, the likelihood of the model cannot be analytically expressed as a function of the marginal log-linear interactions. Therefore at each step of the MCMC algorithm an
iterative procedure is needed in order to calculate the cell probabilities and consequently the model likelihood.

In this paper we construct a novel, fully automatic, efficient MCMC strategy for quantitative learning for graphical log-linear marginal models that handles the problems previously discussed. We assign a suitable prior distribution on the marginal log-linear parameter vector, while the proposal is expressed in terms of the probability parameters. The corresponding proposal on the marginal log-linear parameters is calculated using simple random variable transformation. The advantages of this strategy are clear: the joint distribution factorise under certain conditional independence models, and the likelihood can be directly expressed in terms of probability parameters. In order to construct efficient proposal distributions we use the conditional conjugate approach of Ntzoufras and Tarantola (2013), that exploit the representation of the model in terms of a Direct Acyclic Graph (DAG).

The plan of the paper is as follows. In Section 2, we introduce discrete graphical models of marginal independence and the marginal log-linear parameterisation. In Section 3, we describe the considered prior set-up. Section 4 is devoted to the proposed MCMC strategies. The methodology is illustrated in Section 5 which presents the analysis of Torus Mandibularus dataset. Finally, in Section 6, we conclude with a brief discussion and ideas for future research.

2 Model Specification and Parameterisation

In this section we briefly introduce discrete graphical models of marginal independence, the related notation and terminology, and the corresponding marginal log-linear parameterisation.

A bi-directed graph $G = (V, E)$, is a graph with vertex set $V$, and edge set $E$, such that $(v_i, v_j) \in E$ if and only if $(v_j, v_i) \in E$. Following Richardson (2003) edges are represented via bi-directed arrows. An alternative representation, proposed by Cox and Wermuth (1993), is by undirected dashed edges.

We consider a set of random variables $X_V = (X_v, v \in V)$, each one taking values $i_v \in I_v$; where $I_v$ is the set of possible levels for variable $v$. The cross-tabulation of variables $X_V$ produces a $|V|$-way contingency table with cell frequencies $n = (n(i), i \in I)$ where $I = \times_{v \in V} I_v$. We further assume that $n \sim Multinomial(p, N)$ with $p = \left(p(i), i \in I\right)$; $p(i)$ is the joint probability for cell $i \in I$, and $N = \sum_{i \in I} n(i)$.

A bi-directed graph $G$ is used to represent marginal independencies between variables $X_V$ which are expressed as non-linear constraints over the set of the joint probabilities $p$. The list of independencies implied by a bi-directed graph can be obtained using the pairwise Markov property (Cox and Wermuth, 1993) and the connected set Markov property (Richardson, 2003). For discrete variables the connected set Markov property implies the pairwise Markov property, whereas the converse is not generally true.
Following Drton and Richardson (2008), we define a discrete graphical model of marginal independence as the family of probability distributions for $X_V$ that satisfy the connected set Markov property.

The marginal log-linear parameterisation for bi-directed graphs has been proposed by Lupparelli (2006) and Lupparelli et al. (2009); it is based on the class of log-linear marginal models of Bergsma and Rudas (2002). According to Bergsma and Rudas (2002) the parameter vector $\lambda$, containing the marginal log-linear interactions, can be obtained as

$$\lambda = C \log \left( MP \right) \text{ with } P = \text{vec}(p),$$

where $\text{vec}(p)$ is a vector of dimension $|I|$ obtained by rearranging the elements $p$ in a reverse lexicographical ordering of the corresponding variable levels, with the level of the first variable changing first. Each marginal log-linear interaction satisfies identifiability constraints (here sum-to-zero constraints), and this is achieved via an appropriate contrast matrix $C$. Each interaction is calculated from a specific marginal table identified via the marginalization matrix $M$. More precisely, $M$ specifies from which marginal each element of $\lambda$ is calculated. Each interaction is described by two sets of the variables, one set that refers to the marginal table in use and a second set (a subset of the first one) that identifies which variables are involved in this specific interaction. Finally, the first order interactions correspond to the main effects. Details for the construction of $C$ and $M$ are available in Appendices B and C.

A graphical model of marginal independence is defined by zero constraints on specific marginal log-linear interactions. More precisely, we apply the following procedure presented by Lupparelli (2006) and Lupparelli et al. (2009): (i) define a hierarchical ordering (see Bergsma and Rudas (2002)) of the marginals corresponding to disconnected sets of the bi-directed graph; (ii) append the marginal corresponding to the full table at the end of the list if it is not already included; (iii) for every marginal table estimate all interactions that have not been already obtained from the marginals preceding it in the ordering; (iv) for every marginal table corresponding to a disconnected set of $G$, restrict the highest order log-linear interaction parameter to zero. Hence, the graphical structure imposes constraints of the type

$$K \log \left( MP \right) = 0$$

with $K$ being the sub-matrix of $C$ for which the corresponding elements of $\lambda$ are restricted to zero.

Note that, this parameterisation depends on the ordering of the marginals selected in step (i). Furthermore, it does not always satisfy variation independence; see Lupparelli et al. (2009), Rudas et al. (2010), and Evans and Richardson (2013). If the marginal selected in step (i) are order decomposable then variation independence is guaranteed (Bergsma and Rudas, 2002).
3 Bayesian Model Set-up

3.1 Prior Specification for Marginal Log-linear Interactions

The model in equation (1) can be rewritten in the following extended form

\[
\begin{pmatrix}
\lambda^{M_1} \\
\vdots \\
\lambda^{M_r} \\
\vdots \\
\lambda^{|\mathcal{M}|}
\end{pmatrix} = \text{diag}(C_1, \ldots, C_1, \ldots, C_{|\mathcal{M}|}) \begin{pmatrix}
\log P^{M_1} \\
\vdots \\
\log P^{M_r} \\
\vdots \\
\log P^{|\mathcal{M}|}
\end{pmatrix}, \tag{2}
\]

where \(\mathcal{M} = \{M_1, \ldots, M_r, \ldots, M_{|\mathcal{M}|}\}\) is the set of marginals under consideration, \(\lambda^{M_i}\) is the parameter vector obtained from the marginal probability table \(P^{M_i}\) which is re-arranged to a vector denoted by \(P^{M_i}\).

The contrast matrix \(C\) is a block diagonal matrix with elements \(C_i\). Each sub-matrix \(C_i\) is obtained by inverting the design matrix \(X_{M_i}\) of the saturated model fitted on marginal \(M_i\), and deleting rows corresponding to interactions that are not estimated from that specific marginal table; see Appendix C for the details. From (2), we directly obtain that

\[\lambda^{M_i} = C_i \log P^{M_i}\] for all \(M_i \in \mathcal{M}\).

Every \(\lambda^{M_i}\) may contain interactions that are constrained to zero due the graphical structure \(G\) and the induced contrast matrix. In the following we focus only on non-zero elements of \(\lambda\), on which we assign a suitable prior distribution. We denote by \(\tilde{\lambda}\) the set of elements of \(\lambda\) not restricted to zero, that is

\[\tilde{\lambda} = \left(\tilde{\lambda}^{M_i}; M_i \in \mathcal{M}\right)\] with \(\tilde{\lambda}^{M_i} = \left(\lambda^{M_i}_{\mathbf{j}}; \lambda^{M_i}_{\mathbf{j}} \neq 0, \mathbf{j} = 1, \ldots, r_{C_i}\right)\),

where \(r_{C_i}\) is the number of rows of the contrast matrix \(C_i\) for marginal \(M_i \in \mathcal{M}\).

When no information is available about \(\tilde{\lambda}\) we can simply assign independent normal prior distributions for each element of \(\tilde{\lambda}\) with large variance to express ignorance, i.e.

\[f(\tilde{\lambda}_j) \sim N(0, \sigma_j^2)\] for \(j = 1, 2, \ldots, ||\lambda||_0\),

where \(||\lambda||_0\) is the number of non-zero elements of \(\lambda\).

A more sophisticated approach can be based on the prior suggestion of Dellaportas and Forster (1999) for standard log-linear models; this prior is related to an earlier work by Knuiman and Speed (1988). Dellaportas and Forster (1999) argued that a default prior for the parameters \(\beta\) of a log-linear
model of the form \( \log \mu = X\beta \), where \( \mu \) is the mean vector and \( X \) is a design matrix, is given by

\[
\beta \sim N\left( \theta, 2|I|(X^T X)^{-1} \right),
\]

(3)

where \( |I| = \prod_{v \in V} |I_v| \) is the number of cells of the contingency table and \( \theta \) is a vector of all zeros except the first element that is equal to the logarithm of the average number of observations per cell. Under sum-to-zero constraints, \( (X^T X) \) is a block diagonal matrix resulting to a set of independent priors for interactions referring to different set of variables.

In order to construct the prior distribution on \( \vec{\lambda} \) we work independently on each single set \( \lambda^m \) and we proceed as follows. Let \( \lambda^M \) be the parameter vector for the saturated model that can be estimated from marginal \( M_i \). We can easily find the following connection between our parameterisation and the one by Dellaportas and Forster (1999) \( \lambda^M = \beta^M - \log(N) X^{-1}_M 1 \). Hence, from (3), we get as prior default prior for \( \lambda^M \)

\[
\lambda^M \sim N\left( \theta - \log(N) X^{-1}_M 1, 2|I_M| \left( X^T_M X_M \right)^{-1} \right).
\]

(4)

Finally, since \( \lambda^M \) is a subset of \( \lambda^M \) its prior distribution is obtained from the corresponding marginal distribution of (4). Under sum-to-zero constraints, \( X^{-1}_M 1 \) is a vector of zeros except for the first element that is equal to one. In this case only the prior mean of the intercept differs by a term equal to \( \log(N) \) from the prior of the corresponding element of \( \beta \). If we further assume binary variables and \( \theta = 0 \), then (4) is simplified to a product of independent \( N(0, 2) \) priors for all log-linear parameters.

3.2 Likelihood Specification and Posterior Inference

The likelihood cannot directly be expressed in terms of \( \lambda \) (or equivalently \( \vec{\lambda} \)) but only as a function of the probability parameter:

\[
f(n|\lambda) = \frac{\Gamma(N + 1)}{\prod_{i \in I} \Gamma(n(i) + 1)} \prod_{i \in I} \varphi_i(\lambda)^{n(i)}
\]

where \( \varphi_i(\lambda) \equiv \left\{ p(i) : \lambda = C \log \left( MP \right) \right\} \), for all \( i \in I \)

(5)

Unfortunately, in order to obtain \( \varphi_i(\lambda) \), or equivalently \( P \), from (1), we need to implement an iterative procedure, and therefore the likelihood cannot be written in a closed form expression, see, for example, Bergsma and Rudas (2002), and Lupparelli et al. (2009). As a consequence of this, the corresponding posterior distribution of \( \lambda \) (or equivalently \( \vec{\lambda} \)), cannot be evaluated straightaway. Hence, MCMC methods are needed for posterior inference of \( \vec{\lambda} \).

Although simple Metropolis-Hastings schemes can be used to estimate the posterior distributions of the parameter vector, they will be quite inefficient. In fact, in every MCMC iteration the joint probabilities corresponding to the proposed values of \( \lambda \) need to be calculated using an iterative algorithm,
and this will slow down the MCMC sampler and increase the autocorrelation. Moreover, there could be problems when calculating the joint probability parameters especially if the proposed set of log-linear interactions are not variation independent. If variation independence is not satisfied the resulting set of probabilities are not compatible, i.e. the obtained probabilities may not add to one or be outside the zero-one interval (Qaqish and Ivanova, 2006). For these reasons in the following section we propose alternative MCMC strategy based on the probability representation of the model.

For comparative purposes we have implemented a “vanilla” random walk algorithm MCMC (RWMC) on $\tilde{\lambda}$ which proposes to change each log-linear parameter vector independently for every single marginal. Nevertheless, this algorithm does not ensures that we move within the space of variation independent parametrizations, consequently a well defined joint probability distribution is not always obtained. This problem can be avoided by imposing, at each step of the random walk algorithm, cumbersome restrictions on the conditional posterior distributions that will considerably delay the MCMC sampler. Moreover, we expect that the vanilla RWMC will lead to higher posterior variability in comparison to our proposed method since the algorithm moves to a larger posterior space.

4 Probability Based MCMC Samplers

4.1 Initial Set-up and Data Augmentation for MCMC

Following the notation of Ntzoufras and Tarantola (2013), we can divide the class of graphical log-linear marginal models in two major categories: homogeneous and non-homogeneous models. Both of them can be described by a DAG Markov equivalent over the observed margins. Nevertheless, while homogeneous models can be represented via a DAG with the same vertex set, non-homogeneous ones requires the inclusion of some additional latent variables. The advantage of the DAG representation is that the joint probability over the augmented variable space (including both observed and latent variables) can be written using the standard factorisation. Furthermore, since we work with probabilities in DAGs, the obtained parameterisation is always variation independent (in the simplex). The model can then be parameterised in terms of a minimal set of marginal and conditional probability parameters denoted here with $\Pi$ ($\PiD$ in the original paper) on which we can implement conjugate analysis based on products of Dirichlet distributions.

Using the methodology of Ntzoufras and Tarantola (2013), we construct an MCMC sampler based on proposing values in terms of joint probabilities, avoiding compatibility problems. If the model is homogeneous the dimension of $\Pi$ is the same as the dimension of $\tilde{\lambda}$. This is not true for non-homogeneous models since the dimension of $\Pi$ is greater than the dimension of $\tilde{\lambda}$. In this case we need to augment
the parameter space in order to implement Metropolis-Hastings algorithm.

In the following we denote with $\lambda^A$ the augmented set of marginal log-linear interactions; $\lambda^A = (\tilde{\lambda}, \xi)$ with $\xi$ having dimension equal to $\dim(\Pi) - \dim(\tilde{\lambda})$. If the model is homogeneous $\lambda^A = \tilde{\lambda}$.

More precisely, for any graphical log-linear marginal model $G$ we can obtain a Markov equivalent DAG over the observed margins, denoted by $D_G$, with augmented vertex set $A = \{V, \mathcal{L}\}$, where $\mathcal{L}$ is the set of additional latent variables; if $G$ is homogeneous then $\mathcal{L} = \emptyset$ and $A = V$. Under this approach, the joint probabilities can be written as

$$p(i) = \sum_{i \in I_V} p^A(i, i_L) \text{ for } i \in I_V$$

$$p^A(i, i_L) = p^A(i) = \prod_{v \in A} \pi_{v|pa(v)}(t_v|t_{pa(v)}) \text{ for } i = (i, i_L) \in I_A,$$

and the probability parameter set is given by

$$\Pi = \text{vec}\left(\pi_{v|pa(v)}(t_v|t_{pa(v)}) \text{ for } t_v \in I_V \setminus \{I_v\}, t_{pa(v)} \in I_{pa(v)}, v \in A\right),$$

where $p^A(i) = P(X_V = i_V, X_{\mathcal{L}} = i_{\mathcal{L}})$ is the joint probability for the observed variables $X_V$ and the latent variables $X_{\mathcal{L}}$. $\pi_{v|pa(v)}(j_v|j_{pa(v)}) = P(X_v = j_v|X_{pa(v)} = j_{pa(v)})$ is the conditional probability of each variable $X_v$ given variables $X_{pa(v)}$ in the parent set $pa(v)$ of $v$. By using the induced augmented likelihood representation, we are able to construct a Gibbs sampler based on the conditional conjugate Dirichlet prior distributions on $\Pi$ (Ntzoufras and Tarantola, 2013).

4.2 The general algorithm

The posterior distribution of the augmented set of marginal log-linear interactions is given by

$$f(\lambda^A|n) \propto f\left(n|\varphi(\lambda)\right)f(\tilde{\lambda})f(\xi),$$

where $f(\xi)$ is a pseudo prior used for the additional parameters.

We consider the Metropolis-Hastings algorithm described by the following steps

1. Propose a new vector $\Pi'$ from $q(\Pi'|\Pi^{(t)})$.
2. From $\Pi'$, calculate the proposed joint probabilities $p'$ (for the observed table) using equations (6) and (7).
3. From $p'$, calculate $\lambda'$ using (1) and then obtain the corresponding non-zero elements $\tilde{\lambda}'$.
4. Set $\xi' = \Pi'_\xi$; where $\Pi'_\xi$ is a pre-specified subset of $\Pi'$ of dimension $\dim(\Pi) - \dim(\tilde{\lambda})$. 

8
5. Accept the proposed move with probability $\alpha = \min(1, A)$ with

$$A = \frac{\frac{f(n|\Pi')f(\lambda', \xi')q(\lambda^{(t)}, \xi^{(t)}|\lambda', \xi')}{f(n|\lambda^{(t)})f(\lambda^{(t)})q(\lambda^{(t)}, \xi')}}{\frac{f(n|\Pi)f(\lambda, \xi)q(\Pi^{(t)}|\Pi)}{f(n|\Pi^{(t)})f(\lambda^{(t)}, \xi)q(\Pi^{(t)}, \lambda^{(t)}, \xi)}} \times \abs{\frac{\mathcal{J}(\Pi^{(t)}, \lambda^{(t)}, \xi)}{\mathcal{J}(\Pi', \lambda', \xi')}}$$

where $\abs{\cdot}$ stands for the absolute value, $\Pi, \lambda = \xi$, and $\mathcal{J}(\Pi, \lambda, \xi)$ is the determinant of the jacobian matrix of the transformation $\Pi = g(\lambda, \xi)$ specified by Equations (6), (7), and (1).

6. If the move is accepted, then set $\Pi^{(t+1)} = \Pi'$, $\lambda^{(t+1)} = \xi'$, and $\lambda^{(t+1)} = \lambda'$ otherwise set $\Pi^{(t+1)} = \Pi^{(t)}$ and $\lambda^{(t+1)} = \lambda^{(t)}$.

The pseudo-parameter vector $\xi$ is used only to retain the dimension balance between the marginal log-linear parameterisation and the probability parametrisation used in Ntzoufras and Tarantola (2013). Furthermore, it is directly matched to specific probability parameters of bi-directed graph $G$. Hence, we can indirectly “eliminate” its effect on the algorithm by assuming that its elements are uniformly distributed in the zero-one interval. Under this view, we set $f(\xi) = I_{\{0 < \xi, < 1\}}$ having as a result the elimination of the ratio $f(\xi')/f(\xi^{(t)})$ from (8). In the following, we consider this choice in order to simplify all proposed algorithms.

An efficient proposal for the Metropolis Hastings scheme described above, can be constructed in the following way

$$q(\Pi'|\Pi^{(t)}) = f_q(\Pi'|n^A) f(n^A|\Pi^{(t)}, n)$$  \hspace{1cm} (9)

where $f(n^A|\Pi, n)$ is the distribution of counts $n^A$ given the observed frequency table $n$ and the probability parameter set $\Pi$ of the augmented table induced by $A$; and $f_q(\Pi|n^A)$ is the conditional posterior distribution of the probability parameter vector $\Pi$ given a proposed set of augmented data $n^A$.

In (9), $f(n^A|\Pi, n)$ is readily available from the model construction and the likelihood representation of the augmented table. We only need to specify $f_q(\Pi'|n^A)$ which is the first component of (9) and it has the form of a posterior when a frequency augmented table is given. For this component, we can exploit the conditional conjugate approach of Ntzoufras and Tarantola (2013). In order to do so, we consider as a “prior” $f_q(\Pi)$ a product of Dirichlet distributions in order to obtain a conjugate “posterior” distribution $f_q(\Pi'|n^A)$. Under this approach, (8) becomes equal to

$$A = \frac{f(n|\Pi')f(\lambda', \xi')f_q(\Pi'|n^A)f(n^A|\Pi, n)}{f(n|\Pi^{(t)})f(\lambda^{(t)}, \xi)q(\Pi^{(t)}|\Pi, n)} \times \abs{\frac{\mathcal{J}(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)})}{\mathcal{J}(\Pi', \lambda', \xi')}}$$

$$= \frac{f(n^A|\Pi')f(\lambda', \xi')f_q(\Pi'|n^A)}{f(n^A|\Pi^{(t)})f(\lambda^{(t)}, \xi)q(\Pi'|n^A)} \times \abs{\frac{\mathcal{J}(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)})}{\mathcal{J}(\Pi', \lambda', \xi')}}$$

$$= \frac{f(n^A|\Pi')f(\lambda', \xi')f_q(\Pi'|n^A)}{f(n^A|\Pi^{(t)})f(\lambda^{(t)}, \xi)q(\Pi'|n^A)} \times \abs{\frac{\mathcal{J}(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)})}{\mathcal{J}(\Pi', \lambda', \xi')}}$$

(10)
4.3 Prior Adjustment Algorithm

The MCMC described in Section 4.2 can be considerably simplified by using the following two-step procedure:

**Step 1:** Run the Gibbs sampler of Ntzoufras and Tarantola (2013)

**Step 2:** Use a sub-sample of the previous output as a proposal in the general Metropolis-Hastings algorithm with acceptance rate (8).

By this way, in Step 1, we obtain a random sample of $\Pi$ from the target posterior. We then use a random sub-sample of it in order to eliminate autocorrelation. Using this sub-sample as the proposed values within the Metropolis-Hasting algorithm is equivalent to using the posterior distribution $f_q(\Pi|n)$ as proposal in (8), that is $q(\Pi'|\Pi^{(t)}) = f_q(\Pi'|n)$. Under this proposal, (8) now simplifies to

$$A = \frac{f(\tilde{\lambda}')f_q(\Pi^{(t)})}{f(\lambda^{(t)})f_q(\Pi')} \times \text{abs} \left( \frac{J(\Pi^{(t)}, \tilde{\lambda}^{(t)}, \xi^{(t)})}{J(\Pi', \tilde{\lambda}', \xi')} \right).$$

(11)

We will refer to this algorithm as the the prior-adjustment algorithm (PAA) due to its characteristic to correct for the differences between the prior distributions used under the two parameterisations.

The “prior” distribution $f_q(\Pi)$ is only used to build the proposal. Therefore, it can be considered as a pseudo-prior. It does not influence the target posterior distribution but only affects the convergence rates of PAA. We can choose the parameters of this pseudo-prior in such a way that (11) is maximized such that an optimal acceptance rate is achieved. When a non-informative prior distribution for $\tilde{\lambda}$ is used, all Dirichlet parameters involved in $f_q(\Pi)$ can be set equal to one. Under this choice, the effect of the pseudo-prior is eliminated from the proposal, leaving the data-likelihood to direct the MCMC algorithm. PAA is less computationally demanding that the single-run MCMC algorithm introduced in Section 4.2, since in we avoid four additional likelihood evaluations at each iteration required in the later.

4.4 The Jacobian

We conclude this section by providing analytical expressions of the Jacobian required in the acceptance probabilities within each MCMC step in Sections 4.2 and 4.3; see Equations 10 and 11. Specifically, the
Jacobian terms are given by \( J(\Pi, \tilde{\lambda}, \xi) = \left| \frac{\partial \Pi}{\partial (\tilde{\lambda}, \xi)} \right| \) and are simplified to

\[
J^{-1} = \left| \frac{\partial (\tilde{\lambda}, \xi)}{\partial \Pi} \right| = \left| \frac{\partial (\tilde{\lambda}, \xi)}{\partial (\Pi, \Pi, \Pi, \xi)} \right| = \left| \begin{array}{cccc}
\frac{\partial \tilde{\lambda}}{\partial \Pi} & \frac{\partial \tilde{\lambda}}{\partial \Pi} \\
\frac{\partial \Pi}{\partial \xi} & \frac{\partial \Pi}{\partial \xi} \\
\frac{\partial \Pi}{\partial \xi} & \frac{\partial \Pi}{\partial \xi} \\
\frac{\partial \Pi}{\partial \xi} & \frac{\partial \Pi}{\partial \xi}
\end{array} \right| = -\left| \frac{\partial \tilde{\lambda}}{\partial \Pi, \xi} \right|
\]

where \( \Pi_i, \zeta \) is obtained from \( \Pi \) excluding the elements of \( \Pi, \xi \).

The elements of the Jacobian matrix are given by

\[
\frac{\partial \lambda_k}{\partial \Pi_j} = \sum_{l=1}^{c_C} \left( C_{kl} \left( \sum_{i=1}^{|I|} M_{li} P_i \right)^{-1} \sum_{i=1}^{|I|} M_{li} \Delta_{ij} \right) \text{ with } \Delta_{ij} = \frac{\partial P_i}{\partial \Pi_j},
\]

where \( P_i \) denote the \( i \) element of \( P \), and \( c_C \) is the number of columns of the contrast matrix \( C \). For the saturated model, the above equation simplifies to

\[
\frac{\partial \lambda_k}{\partial P_j} = \sum_{l=1}^{c_C} \frac{C_{kl}(M_{lj} - M_{l|I|})}{\sum_{i=1}^{|I|} M_{li} P_i}
\]

since \( \Pi = P \), and \( p_{|I|} = 1 - \sum_{i=1}^{|I|} P_i \). More details on the calculation of (12) are provided in Appendix A.

In order to complete the specification of (12), we need to calculate the derivative terms \( \Delta_{ij} \). Let us now denote by \( i \) every cell in \( I \) that corresponds to index \( i \) and therefore \( P_i = p(i) \). Similarly, the index \( j \) will define a specific cell \( j \in I^A \) and a variable \( u_j \in A \) such that \( \Pi_j = \pi_{u_j|pa(u_j)}(j_{u_j|j_{pa(u)}}) \). Therefore, terms \( \Delta_{ij} \) can be written as

\[
\Delta_{ij} = \frac{\partial p(i)}{\partial \pi_{u_j|pa(u_j)}(j_{u_j|j_{pa(u)}})} \text{ for every } i \mapsto i \text{ and } j \mapsto (u_j, j).
\]

For the computation of \( \Delta_{ij} \) we consider two different cases: (A) \( u_j \in \mathcal{V} \) and (B) \( u_j \in \mathcal{L} \). In the following, to simplify notation, we denote \( u_j \) by \( u \).

For case A, when \( u \) is an observed variable, we obtain

\[
\frac{\partial p(i)}{\partial \pi_{u_j|pa(u)}(j_{u_j|j_{pa(u)}})} = \delta(i, j) \frac{p_A(i, j_{L})}{\pi_{u_j|pa(u)}(j_{u_j|j_{pa(u)}})}
\]

with

\[
\delta(i, j) = \begin{cases} 
0 & \text{if } j_u \neq i_u \text{ and } i_{pa(u)} \neq j_{pa(u)} \text{ or } i_{pa(u)} \neq j_{pa(u)} \text{ or } i_{pa(u)} \neq j_{pa(u)} \\
1 & \text{if } j_u = i_u \text{ and } i_{pa(u)} \neq j_{pa(u)} \\
-1 & \text{if } j_u = i_u \text{ and } i_{pa(u)} = j_{pa(u)} \text{ and } i_{pa(u)} \neq j_{pa(u)} \text{ or } i_{pa(u)} \neq j_{pa(u)} \text{ or } j_{pa(u)} \neq j_{pa(u)} \text{ or } j_{pa(u)} \neq j_{pa(u)} 
\end{cases}
\]
where $\mathcal{L}_u = \mathcal{L} \cap \text{pa}(u)$, $\mathcal{A}_u = \mathcal{V} \cup \{u\} \cup \mathcal{L}_u$, and

$$p^{A_u}(i, j|\mathcal{L}_u) = \begin{cases} P(X_{\mathcal{V}} = i, X_{\mathcal{L} \cap \text{pa}(u)} = j|\mathcal{L} \cap \text{pa}(u)) & \mathcal{L}_u \neq \emptyset \\ p(i) & \mathcal{L}_u = \emptyset \end{cases}.$$  

For **case B**, when $u$ is a latent variable, then $u$ does not have any parents ($\text{pa}(u) = \emptyset$) due to the structure of the DAG representation. Hence, the derivative is given by

$$\frac{\partial p(i)}{\partial \pi_u|\text{pa}(u)}(j_u|j_{\text{pa}(u)}) = \frac{\partial p(i)}{\partial \pi_u}(j_u) = \frac{p^{A_u}(i, j_u)}{\pi_u(j_u)} - \frac{p^{A_u}(i, |\mathcal{I}_u|)}{\pi_u(|\mathcal{I}_u|)} \text{ for } j_u < |\mathcal{I}_u|. \tag{16}$$

Detailed derivation of expressions (14)–(16) are available in Appendix A.

### 5 Illustrative Example

We illustrate the proposed methodology by using the dataset of Muller and Mayhall (1971) studying the incidence of the morphological trait torus mandibularis in different Eskimo groups. Torus mandibularis is a bony growth in the mandible along the surface nearest to the tongue. This morphological structure of the mouth is frequently used by anthropologist to study differences among populations and among groups within the same population. This data have been previously analysed by Bishop *et al.* (1975) via log linear models, and by Lupparelli (2006) via marginal log-linear graphical models.

For our analysis, we consider the data presented in Table 1, cross-classifying age (A), incidence of Torus Mandibularis (I), sex (S) and population (P). The dataset is a dichotomized version of the original data of Muller and Mayhall (1971). The examined Eskimo groups refers to different geographical regions, Igloolik and Hall Beach groups are from Foxe Basin area of Canada whereas Aleut are from Western Alaska. Furthermore, the data of the Aleuts group were collected by an investigator different from the one who collected the data for the first two groups, with a time difference between investigations of about twenty years. For the previous reasons we decided to reclassify the data in two groups: the first one including Igloolik and Hall Beach and the second one Aleut. Finally, variable age has been classified in two groups according to the median value.

From the analysis of Lupparelli (2006) we know that model SP+PI+IA fits well the original data, hence, in the following, we concentrate on this specific four-chain graph.

Table 2 reports the posterior means and standard deviations for the marginal log-linear interactions obtained via 10000 iteration with a burn in of 1000 with the proposed PAA algorithm and the RWMCMC. The maximum likelihood estimates (MLEs) and corresponding approximate standard errors are also reported for comparative purposes. Moreover, Figure 1 depicts the posterior ergodic plots for the estimates of the elements of $\hat{X}$ obtained using PAA and the corresponding approximate MLE based statistics.
Figure 1: Ergodic plots for marginal log-linear interactions using Prior-Adjustment Algorithm (PAA) compared with approximate maximum likelihood estimates for the Torus Mandibularis data.
Table 1: Torus Mandibularis in Eskimo Populations

<table>
<thead>
<tr>
<th>Population (P)</th>
<th>Sex (S)</th>
<th>Incidence (I)</th>
<th>Age Groups (A)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1-20</td>
<td>Over 20</td>
</tr>
<tr>
<td>Igloolik and Hall Beach</td>
<td>Male</td>
<td>Present</td>
<td>19</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absent</td>
<td>103</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Present</td>
<td>16</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absent</td>
<td>87</td>
<td>36</td>
</tr>
<tr>
<td>Aleut</td>
<td>Male</td>
<td>Absent</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Present</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Present</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absent</td>
<td>17</td>
<td>20</td>
</tr>
</tbody>
</table>

From Table 2 and Figure 1 we observe that the posterior estimates (for both MCMC methods) and the MLEs coincide for all interactions and main effects obtained by marginals where no latent variable is involved. This is not the case for $\lambda_{AISP}(2, 2)$, $\lambda_{AISP}(2, 2, 2)$, $\lambda_{AISP}(2, 2)$, where the latent variable is involved. More specifically, the posterior standard deviations are lower by 6.5%, 34% and 26%, respectively. This result is intuitively expected since PAA moves across the correct posterior distribution of variation independent parameterisation. On the other hand, while both the RWMCMC and the approximate MLEs standard errors are obtained without considering the restrictions imposed in order to obtain a variation independent parameterisation. Finally, differences are also observed for interaction $\lambda_{AISP}(2, 2, 2)$ where RWMCMC provides posterior means far away from the corresponding MLEs with PAA being quite closely and standard deviance slightly higher than both the corresponding values of RWMCMC and the MLEs standard errors.

In terms of computational time, the PAA was found to be faster by 10% than the RWMCMC (126.5 versus 140.5 seconds for 1000 iterations).

The effective sample size (ESS) is higher for the parameters where latent specification is involved (i.e. IP, AIP, ISP, AISP) and for the main effect A with the increase ranging from 13% to 49%. For the remaining parameters parameters the ESS for our method was lower with values ranging from 7% to 37%.

Taking into consideration also the computational time of the two algorithms, the MCMC efficiency (ESS/time) is even more improved for our proposed algorithm with the ESS per second being higher
with the increase ranging from 26% to 63% for the parameters discussed earlier. The average relative MCMC efficiency is higher for our method than the random walk MCMC by 11%. The above statistics were calculated with the effectiveSize function of coda package in R. The overall picture is similar if we consider results using the monitor function of rstan package in R.

Finally, the MCMC errors were lower for the majority of the parameters by using the naive estimator of coda package with relative values ranging from 0.66 up to 1.06 while the corresponding relative values for the time series based estimator are ranging from 0.61 up to 1.31.

### 6 Discussion

A possible way to parametrise discrete graphical models of marginal independence is by using the log-linear marginal models of Bergsma and Rudas (2002). The marginal log-linear interactions are calculated

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PAA Mean</th>
<th>PAA SD</th>
<th>RWCMC Mean</th>
<th>RWCMC SD</th>
<th>ML Estimate</th>
<th>ML SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_0^{AS}$</td>
<td>-1.391</td>
<td>0.004</td>
<td>-1.391</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_1^{AS}(2)$</td>
<td>-0.001</td>
<td>0.042</td>
<td>-0.003</td>
<td>0.043</td>
<td>-0.002</td>
<td>0.043</td>
</tr>
<tr>
<td>$\lambda_2^{AS}(2)$</td>
<td>-0.072</td>
<td>0.043</td>
<td>-0.079</td>
<td>0.043</td>
<td>-0.072</td>
<td>0.043</td>
</tr>
<tr>
<td>$\lambda_3^{AS}(2,2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_4^{AP}(2)$</td>
<td>-0.697</td>
<td>0.053</td>
<td>-0.695</td>
<td>0.055</td>
<td>-0.699</td>
<td>0.054</td>
</tr>
<tr>
<td>$\lambda_5^{AP}(2,2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_6^{IS}(2)$</td>
<td>0.234</td>
<td>0.045</td>
<td>0.241</td>
<td>0.044</td>
<td>0.232</td>
<td>0.044</td>
</tr>
<tr>
<td>$\lambda_7^{IS}(2,2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_8^{ISP}(2,2)$</td>
<td>0.004</td>
<td>0.053</td>
<td>-0.009</td>
<td>0.055</td>
<td>0.003</td>
<td>0.054</td>
</tr>
<tr>
<td>$\lambda_9^{ISP}(2,2,2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_{10}^{AIS}(2,2)$</td>
<td>-0.509</td>
<td>0.051</td>
<td>-0.505</td>
<td>0.052</td>
<td>-0.507</td>
<td>0.051</td>
</tr>
<tr>
<td>$\lambda_{11}^{AIS}(2,2,2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$\lambda_{12}^{AISP}(2,2)$</td>
<td>0.057</td>
<td>0.058</td>
<td>0.082</td>
<td>0.063</td>
<td>0.052</td>
<td>0.062</td>
</tr>
<tr>
<td>$\lambda_{13}^{AISP}(2,2,2)$</td>
<td>0.132</td>
<td>0.068</td>
<td>0.049</td>
<td>0.065</td>
<td>0.151</td>
<td>0.062</td>
</tr>
<tr>
<td>$\lambda_{14}^{AISP}(2,2,2)$</td>
<td>0.029</td>
<td>0.041</td>
<td>0.066</td>
<td>0.063</td>
<td>0.072</td>
<td>0.062</td>
</tr>
<tr>
<td>$\lambda_{15}^{AISP}(2,2,2)$</td>
<td>0.047</td>
<td>0.046</td>
<td>0.034</td>
<td>0.063</td>
<td>0.037</td>
<td>0.062</td>
</tr>
</tbody>
</table>
from specific marginals of the original table, and independences imply zero constraints on specific set of interactions in a similar manner as in conditional log-linear graphical models.

In this work we focus on the Bayesian estimation of the log-linear parameters for graphical models of marginal independence. Bayesian analysis of such models is not widespread mainly due to the computational problems involved in the derivation of their posterior distribution. More specifically, MCMC methods need to be used since no conjugate analysis is available. Major difficulties arise from the fact that we need to move within a posterior distribution satisfying restrictions implied by variation independence parametrizations. In the proposed algorithm, we satisfy these restrictions by sampling from the probability space of the graphical model of marginal independence under consideration. Then we transform the probability parameter values to the corresponding marginal log-linear ones avoiding the iterative procedure needed for the evaluation of the likelihood. In order to achieve this, we exploit the augmented DAG representation of the model. This not only facilitates prior elicitation but also the construction of the jacobian matrix involved in the acceptance probability of the induced Metropolis steps. Even if the derivation of the proposed algorithm is elaborate, it leads to an efficient and fully automatic setup. By this way we sample directly from the target posterior, and, on the same time, we avoid any time consuming and troublesome tuning of MCMC parameters.

An alternative approach can be solely based on proposing values for the log-linear parameters directly via the corresponding probability parameters. For example, we may use a Metropolis algorithm with random walk proposals on the logits of each probability. Then (8) becomes equal to

$$A = \frac{f(n|\Pi')f(\lambda')}{f(n|\Pi^{(t)})f(\lambda^{(t)})} \times \frac{\Pi'(1 - \Pi')}{\Pi^{(t)}(1 - \Pi^{(t)})} \times \text{abs} \left( \frac{\mathcal{J}(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)})}{\mathcal{J}(\Pi', \lambda', \xi')} \right)$$

(17)

By this way, we may avoid dependencies on the conditional conjugate analysis used in PAA or the generation of additional latent data as in the general framework of Section 4.

Nevertheless, we will not any more have a fully automatic method since we will need to tune the variances $\sigma_i^2$ of the proposal distributions. This MCMC version may be preferable when the number of latent variables introduced in the DAG representation $D_G$ is large. Although we initially experimented with this approach, we decided to focus on “automatic” setups proposed in this article since they were more efficient in practice.

For the future, the authors would like to exploit and study the connections between the prior and the posterior distributions for the two different parameterisations (probability versus marginal log-linear). Moreover, extension of the method to accommodate fully automatic selection, comparison and model averaging techniques is an intriguing topic for further investigation.
Supplementary Material


Acknowledgments

We would like to thank Giovanni Marchetti for providing us the R function inv.mlogit. This research was partially funded by the Research Centre of the Athens University of Economics and Business (Funding program for the research publications of the AUEB Faculty members) and by the Department of Economics and Management of University of Pavia.

References


