Bayesian Analysis of Graphical Models of Marginal Independence for Three Way Contingency Tables

Ioannis Ntzoufras *

Department of Statistics, Athens University of Economics and Business, Greece

Claudia Tarantola

Department of Economics and Quantitative Methods, University of Pavia, Italy

October 25, 2010

Abstract

This paper deals with the Bayesian analysis of graphical models of marginal independence for three way contingency tables. Each marginal independence model corresponds to a particular factorization of the cell probabilities and a conjugate analysis based on Dirichlet prior can be performed. We illustrate a comprehensive Bayesian analysis of such models, involving suitable choices of prior parameters, estimation, model determination, as well as the allied computational issues. The posterior distributions of the marginal log-linear parameters is indirectly obtained using simple Monte Carlo schemes. The methodology is illustrated using two real data sets.

*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 parameterization, Monte Carlo computation, order decomposability, power prior approach.

1 Introduction

The use of graphical models to describe association between categorical variables dates back to the work of Darroch *et al.* (1980), where graphical log-linear models were introduced. Since this initial work, graphical models turns out to be an efficient methodology for categorical data analysis. Different typology of graphs have been proposed and the corresponding model selection methodology developed. In this paper we focus of on graphical models of marginal independence, see Cox and Wermuth (1993).

Graphical models of marginal independence were originally introduced for the analysis of multivariate Gaussian distributions. They compose a family of multivariate distributions incorporating the marginal independences represented by a bi-directed graph. The nodes in the graph correspond to a set of random variables and the bi-directed edges represent the pairwise associations between them. A missing edge from a pair of nodes indicates that the corresponding variables are marginally independent.

The analysis of the Gaussian case can be easily performed both in classical and Bayesian frameworks since marginal independences correspond to zero constraints in the variance-covariance matrix. The situation is more complicated in the discrete case, where marginal independences correspond to non linear constraints on the set of parameters. Only recently parameterizations for these models have been proposed by Lupparelli (2006), Lupparelli *et al.* (2008) and Drton and Richardson (2008).

In this paper we focus on the analysis of three way contingency tables. In the three

way case, the joint probability of each model under consideration can be appropriately factorized, hence we can work directly in terms of the vector of joint probabilities on which we impose the constraints implied by the graph. We consider a minimal set of probability parameters expressing marginal/conditional independences and sufficiently describe the graphical model of interest. We use a conjugate prior distribution based on Dirichlet priors on the appropriate probability parameters. In order to make the prior distributions 'compatible' across models we define all probability parameters (marginal and conditional ones) of each model from the parameters of the joint distribution of the full table. The prior parameters of the Dirichlet distribution for the saturated model in the full table are specified using ideas based on the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000). We discuss the effects of different choices of the hyper-parameter values on the model selection results. Finally, the posterior distributions of the corresponding marginal log-linear parameters are obtained via Monte Carlo simulations.

The plan of the paper is as follows. In Section 2 we introduce graphical models of marginal independence, we establish the notation and we present their global Markov property. In Section 3 we present two possible parameterizations and illustrates a suitable factorization of the likelihood function applicable if the probability parameterizations is used. In Section 4, we consider conjugate prior distributions, we present an imaginary data approach for prior specification and we compare alternative prior setups. Section 5 provides posterior model and parameter distributions. Two illustrative examples are presented in Section 6. Finally, we end up with a discussion and some final comments regarding our current research on the topic.

2 Graphical Models of Marginal Independence

In this section we briefly introduce graphical models of marginal independence, for more details see e.g. Drton and Richardson (2008).

A bi-directed graph $G = (\mathcal{V}, E)$ is characterized by a vertex set $\mathcal{V}$ and an edge set E with the property that $(v_i, v_j) \in E$ if and only if $(v_j, v_i) \in E$. We denote each bi-directed edge by $(\overleftrightarrow{v_i, v_j}) = \big\{(v_i, v_j), (v_j, v_i)\big\}$ and we represent it with a bi-directed arrow. A path connecting two vertices, v_0 and v_m, is a finite sequence of distinct vertices $v_0, \dots v_m$ such that $(v_{i-1}, v_i),\ i = 1, \dots, m$, is an edge of the graph. A vertex set $C \subseteq \mathcal{V}$ is connected if every two vertexes v_i and v_j are joined by a path in which every vertex is in C. Two sets $S_1, S_2 \in \mathcal{V}$ are separated by a third set $S_3 \in \mathcal{V}$ if any path from a vertex in S_1 to a vertex in S_2 contains a vertex in S_3. It can be shown that, if a subset of the nodes D is not connected then there exist a unique partition of it into maximal (with respect to inclusion) connected set $C_1, \dots, C_r$

$$D = C_1 \cup C_2 \cup \dots \cup C_r. \tag{1}$$

The graph is used to represent marginal independences between a set of discrete random variables $X_{\mathcal{V}} = \big(X_v,\ v \in \mathcal{V}\big)$, 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 contingency table of dimension $|\mathcal{V}|$ with cell frequencies $\boldsymbol{n} = \Big(n(i),\ i \in \mathcal{I}\Big)$ where $\mathcal{I} = \bigtimes_{v \in \mathcal{V}} \mathcal{I}_v$. Similarly for any marginal $M \subseteq \mathcal{V}$, we denote with $X_M = \big(X_v,\ v \in M\big)$ the set of variables which produce the marginal table with frequencies $\boldsymbol{n}_M = \Big(n_M(i_M),\ i_M \in \mathcal{I}_M\Big)$ where $\mathcal{I}_M = \bigtimes_{v \in M} \mathcal{I}_v$. The marginal cell counts are the sum of

specific elements of the full table and are given by

$$n_M(i_M) = \sum_{j_{\overline{M}} \in \mathcal{I}_{\overline{M}}} n(i_M, j_{\overline{M}}) \; = \sum_{j \in \mathcal{I}: j_M = i_M} n(j)$$

where $\overline{M} = \mathcal{V} \setminus M$. For the three way case, we use the simplified notation with $n(i)$ denoted by n_{abc} for every $i = (i_A = a, i_B = b, i_C = c)$ where A, B and C denote the three nodes of the graph. The corresponding counts $n_M(i_M)$ of the marginal M will be denoted by putting the plus $(+)$ sign on the index of every variable $v \in \overline{M}$. For example the $n_B(i_B)$ (i.e. $M = \{B\}$) will be denoted by n_{+b+} while the marginal $n_{AC}(i_{AC})$ (i.e. $M = \{A, C\}$) will be denoted by n_{a+c}. Similar notation will be also used for the cell probabilities.

The list of independences implied by a bi-directed graph can be obtained using the global Markov property (Kauermann, 1996 and Richardson, 2003). The distribution of a random vector $X_{\mathcal{V}} = \{X_v, v \in V\}$ satisfies the global Markov property if

$$S_1 \text{ is separated from } S_2 \text{ by } V \setminus (S_1 \cup S_2 \cup S_3) \text{ in } G \text{ implies } X_{S_1} \perp\!\!\!\perp X_{S_2} | X_{S_3}, \qquad (2)$$

with S_1, S_2 and S_3 disjoint subsets of V, and S_3 may be empty.

From the global Markov property, we directly derive that if two nodes i and j are disconnected then $X_i \perp\!\!\!\perp X_j$, that is the variables are marginal independent. This can be easily generalized for any given disconnected set D satisfying (1). Then the global Markov property for the bi-directed graph G implies that $X_{C_1} \perp\!\!\!\perp X_{C_2} \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_{C_r}$..

The previous property can de used to define a bi-directed discrete graphical model. According to Drton and Richardson (2008), a discrete graphical model of marginal independence associated to a bi-directed graph G is a family $P(G)$ of joint distributions for a categorical random vector $X_{\mathcal{V}}$ satisfying the global Markov property. Following

the above, for every not connected set $D \subseteq \mathcal{V}$, it holds that

$$P(X_D = i_D) = \prod_{k=1}^{r} P(X_{C_k} = i_{C_k}) \tag{3}$$

where $C_1, \ldots, C_r$ are the inclusion maximal connected sets satisfying (1).

3 Parameterizations for discrete graphical models of marginal independence

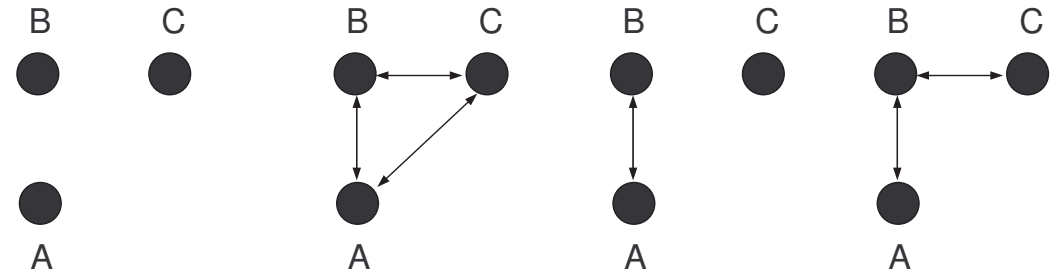
A graphical model of marginal independence for a three way contingency table can be parameterized both in terms of the cell-probabilities (which for simplicity we call the π-parameterization), or in terms of the marginal log-linear parameterization of Lupparelli *et al.* (2009) (named the λ-parameterization).

3.1 Probability parameterization and likelihood factorization

Under the π-parameterization we impose the constraints implied by the graph G directly on the joint probabilities $\boldsymbol{\pi}$. We work with a minimal set of probability parameters $\boldsymbol{\pi}^G$ expressing marginal/conditional independences and sufficiently describe the graphical model G under investigation.

In the three way case the joint probability of each model can be appropriately factorized for any graph G. For every three way contingency table eight possible graphical models exist which can be represented by four different types of graphs: the independence, the saturated, the edge (only one edge) and the gamma structure graph (a single path of length two). The different types of graph are represented in Figure 1.

For the saturated model G_S, we get all parameters from the full table, i.e. $\boldsymbol{\pi}^{G_S} = \boldsymbol{\pi} = \big(\pi(i),\, i \in \mathcal{I}\big) = \big(\pi_{abc},\, a = 1, \ldots, |\mathcal{I}_A|, b = 1, \ldots, |\mathcal{I}_B|, c = 1, \ldots, |\mathcal{I}_C|\ \big)$. Thus, the



(a) Independence Model (b) Saturated Model (c) Edge Model (d) Gamma Model

Figure 1: Type of Graphs in Three Way Tables

likelihood is

$$f(\boldsymbol{n}|\boldsymbol{\pi}, G_S) = \mathsf{C}(\boldsymbol{n}) \prod_{i\in\mathcal{I}} \pi(i)^{n(i)} = \mathsf{C}(\boldsymbol{n}) \prod_{a=1}^{|\mathcal{I}_A|} \prod_{b=1}^{|\mathcal{I}_B|} \prod_{c=1}^{|\mathcal{I}_C|} \pi_{abc}^{n_{abc}}$$

where

$$\mathsf{C}(\boldsymbol{n}) = \frac{\Gamma(N+1)}{\prod_{i\in\mathcal{I}} \Gamma\big(n(i)+1\big)} = \frac{\Gamma(N+1)}{\prod_{a=1}^{|\mathcal{I}_A|} \prod_{b=1}^{|\mathcal{I}_B|} \prod_{c=1}^{|\mathcal{I}_C|} \Gamma\big(n_{abc}+1\big)}$$

with $N = \sum_{i\in\mathcal{I}} n(i) = \sum_{a=1}^{|\mathcal{I}_A|} \sum_{b=1}^{|\mathcal{I}_B|} \sum_{c=1}^{|\mathcal{I}_C|} n_{abc}$ being the total sample size.

The joint distributions for the independence model $\{A, B, C\}$ and the edge models of type $\{e, \overline{e}\}$ can be factorized using (3) since their graphical structure implies that $A \perp\!\!\!\perp B \perp\!\!\!\perp C$ for the first and $e \perp\!\!\!\perp \overline{e}$ for the latter; where $e \in \mathcal{V} = \{A, B, C\}$ is the disconnected variable of the edge graph and $\overline{e} = \mathcal{V} \setminus e$. For example, the edge model $\{AB, C\}$ with $e = C$ implies that $AB \perp\!\!\!\perp C$.

Hence for the independence model we have that $\pi_{abc} = \pi_{a++}\pi_{+b+}\pi_{++c}$ resulting in

$$\begin{aligned} f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) &= \mathsf{C}(\boldsymbol{n}) \prod_{v\in\{A,B,C)} \prod_{i_v\in\mathcal{I}_v} \pi_v(i_v)^{n(i_v)} \\ &= \mathsf{C}(\boldsymbol{n}) \prod_{a=1}^{|\mathcal{I}_A|} (\pi_{a++})^{n_{a++}} \prod_{b=1}^{|\mathcal{I}_B|} (\pi_{+b+})^{n_{+b+}} \prod_{c=1}^{|\mathcal{I}_C|} (\pi_{++c})^{n_{++c}} \end{aligned}$$

where $\boldsymbol{\pi}^G = \big(\boldsymbol{\pi}_A, \boldsymbol{\pi}_B, \boldsymbol{\pi}_C\big)$, $\boldsymbol{\pi}_A = (\pi_{a++}, a = 1, \ldots, |\mathcal{I}_A|)$, $\boldsymbol{\pi}_B = (\pi_{+b+}, b = 1, \ldots, |\mathcal{I}_B|)$ and $\boldsymbol{\pi}_C = (\pi_{++c}, c = 1, \ldots, |\mathcal{I}_C|)$.

The edge graphs are characterized by the disconnected variable e. Three possible graphs/models are included for each choice of $e \in \{A, B, C\}$ which imply the independences $A \perp\!\!\!\perp (B, C)$, $B \perp\!\!\!\perp (A, C)$ and $C \perp\!\!\!\perp (A, B)$ respectively. From (3) we have that $\pi(i) = \pi_e(i_e)\pi_{\overline{e}}(i_{\overline{e}})$ resulting in a likelihood of the following form

$$f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) \;\; = \;\; \mathsf{C}(\boldsymbol{n}) \prod_{i_e \in \mathcal{I}_e} \pi_e(i_e)^{n(i_e)} \prod_{i_{\overline{e}} \in \mathcal{I}_{\overline{e}}} \pi_{\overline{e}}(i_{\overline{e}})^{n(i_{\overline{e}})}$$

with $\boldsymbol{\pi}^G = \big(\boldsymbol{\pi}_e, \boldsymbol{\pi}_{\overline{e}}\big)$. For example if $e = C$ then the likelihood can be rewritten as

$$f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) \;\; = \;\; \mathsf{C}(\boldsymbol{n}) \prod_{b=1}^{|\mathcal{I}_B|} (\pi_{++c})^{n_{++c}} \prod_{a=1}^{|\mathcal{I}_A|} \prod_{b=1}^{|\mathcal{I}_B|} (\pi_{ab+})^{n_{ab+}}$$

with $\boldsymbol{\pi}^G = \big(\boldsymbol{\pi}_{AB}, \boldsymbol{\pi}_C\big)$ and $\boldsymbol{\pi}_{AB} = \big(\pi_{ab+}, a = 1, \ldots, |\mathcal{I}_A|, b = 1, \ldots, |\mathcal{I}_B|\big)$.

Finally, a gamma structured model is characterized by its corner node v_c which is connected with both variables in $\overline{v}_c = \mathcal{V} \setminus v_c$. If we denote the two variables of $\overline{v}_c$ by v_1 and v_2, then a gamma structured model implies that

$$v_1 \text{ is separated from } v_2 \text{ by } v_c.$$

Applying the global Markov property (see (2)) on the above relationship results in $v_1 \perp\!\!\!\perp v_2$ since $v_c = \mathcal{V} \setminus \big(\{v_1\} \cup \{v_2\} \cup \emptyset\big)$ and $\mathcal{V} = v_c \cup \overline{v}_c = \{v_1, v_2, v_c\}$. To implement the above marginal independence, we need to write $\pi(i) = \pi_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c})\pi_{\overline{v}_c}(i_{\overline{v}_c})$ and then substitute $\pi_{\overline{v}_c}(i_{\overline{v}_c})$ by the $\pi_{v_1}(i_{v_1})\pi_{v_2}(i_{v_2})$ resulting in

$$\pi(i) = \pi_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c})\pi_{v_1}(i_{v_1})\pi_{v_2}(i_{v_2}).$$

In the above, $\pi_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c})$ denotes the conditional probability $P(v_c = i_{v_c}|\overline{v}_c = i_{\overline{v}_c})$ for

any $i = \left(i_{v_c}, i_{\overline{v}_c}\right)$. The above parameterization results in the following likelihood

$$f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) = \mathsf{C}(\boldsymbol{n}) \prod_{i_{\overline{v}_c} \in \mathcal{I}_{\overline{v}_c}} \left(\prod_{i_{v_c} \in \mathcal{I}_{v_c}} \pi_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c})^{n(i_{v_c}, i_{\overline{v}_c})} \right) \prod_{i_{v_1} \in \mathcal{I}_{v_1}} \pi_{v_1}(i_{v_1})^{n(i_{v_1})} \prod_{i_{v_2} \in \mathcal{I}_{v_2}} \pi_{v_2}(i_{v_2})^{n(i_{v_2})}$$

where $\boldsymbol{\pi}^G = \left(\boldsymbol{\pi}_{v_c|\overline{v}_c}, \boldsymbol{\pi}_{v_1}, \boldsymbol{\pi}_{v_2}\right)$ and $\boldsymbol{\pi}_{v_c|\overline{v}_c} = \left(\pi_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c}),\ i_{v_c} \in \mathcal{I}_{v_c},\ i_{\overline{v}_c} \in \mathcal{I}_{\overline{v}_c}\right)$ are all the conditional probabilities of v_c given $\overline{v}_c$. For example if $v_c = B$ then the likelihood is written as

$$f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) = \mathsf{C}(\boldsymbol{n}) \prod_{a=1}^{|\mathcal{I}_A|} \prod_{c=1}^{|\mathcal{I}_C|} \left(\prod_{b=1}^{|\mathcal{I}_B|} \left[\pi_{B|AC}(b|ac)\right]^{n_{abc}} \right) \prod_{a=1}^{|\mathcal{I}_A|} (\pi_{a++})^{n_{a++}} \prod_{c=1}^{|\mathcal{I}_C|} (\pi_{++c})^{n_{++c}}$$

where $\boldsymbol{\pi}^G = \left(\boldsymbol{\pi}_{B|AC}, \boldsymbol{\pi}_A, \boldsymbol{\pi}_C\right)$.

3.2 Marginal log-linear parameterization

Log-linear models are widely used to represent the conditional independence relations depicted by an undirected graph but they cannot be used to describe properties of the marginal distributions. These models have been adapted to allow the analysis of marginal contingency tables; see, for example, McCullagh and Nelder (1989), Liang et al. (1992), Lang and Agresti (1994), Glonek and McCullagh (1995), and Bergsma and Rudas (2002). In particular Bergsma and Rudas (2002) introduced marginal log-linear models as a generalization of ordinary log-linear and multivariate logistic models. Marginal log-linear parameters are obtained in a similar fashion as ordinary log-linear parameters but estimated using the frequencies of appropriate marginal contingency tables rather than data of the entire contingency table. A marginal log-linear parameter is described by two sets of the variables: one set that refers to the marginal table in use and a second set (which is a subset of the first) that identifies which variables are involved in this specific term/effect. This setup is important in cases where 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. Marginal log-linear models measure average conditional association among the variables involved. Thus they describe marginal associations in terms of log-odds ratios which can be useful to summarize associations when some discrete covariates are not available.

The marginal log-linear parameters of Bergsma and Rudas (2002) can be obtained by

$$\boldsymbol{\lambda} = \boldsymbol{C} \log \Big(\boldsymbol{M} \mathrm{vec}(\boldsymbol{\pi}) \Big) \tag{4}$$

where $\boldsymbol{\pi} = \Big(\pi(i),\ i \in \mathcal{I} \Big)$ is the joint probability distribution of $X_{\mathcal{V}}$ and $\mathrm{vec}(\boldsymbol{\pi})$ is a vector of dimension $|\mathcal{I}|$ obtained by rearranging the elements $\boldsymbol{\pi}$ in a reverse lexicographical ordering of the corresponding variable levels with the level of the first variable changing first (or faster). In this paper we assume that the parameter vector $\boldsymbol{\lambda}$ satisfies sum-to-zero constraints and we indicate with $\boldsymbol{C}$ the corresponding contrast matrix. Finally $\boldsymbol{M}$ is the marginalization matrix which specifies from which marginal we calculate each element of $\boldsymbol{\lambda}$. Such models have been also used by Lupparelli (2006) and Lupparelli *et al.* (2008) to parameterize marginal association graphs. Each log-linear parameter is calculated within the appropriate marginal distribution and a graphical model of marginal independence is defined by zero constraints on specific higher order marginal log-linear parameters. Following this approach, we can obtain an algorithm for constructing $\boldsymbol{C}$ and $\boldsymbol{M}$ matrices which is available in the Appendix (for additional details see Appendix A in Lupparelli, 2006).

To obtain a λ-parameterization we need to follow the steps described by Lupparelli (2006) and Lupparelli *et al.* (2009): (i) construct any hierarchical ordering (see

Bergsma and Rudas, 2002) of the marginals M_i corresponding to disconnected sets of the graph denoted by $\mathcal{D}(G)$, i.e. for $M_i \in \mathcal{D}(G)$; (ii) append the marginal $M = \mathcal{V}$ (corresponding to the full table under consideration) at the end of the list if it is not already included; (iii) for every marginal table $M_i \in \mathcal{D}(G) \cup \mathcal{V}$ estimate all parameters of effects in M_i that have not already estimated from the preceding marginals; (iv) for every marginal table $M_i \in \mathcal{D}(G)$, set the highest order log-linear interaction parameter equal to zero; see Proposition 4.3.1 in Lupparelli (2006). Following this procedure we can obtain the log-linear parameters for the bi-directed graphs of Figure 1 as reported in Table 1.

The problem with $\lambda-$parameterization is that we cannot use (4) to obtain a closed form expression for $\boldsymbol{\pi}^G$ and $\boldsymbol{\pi}$. Thus the likelihood is not analytically available and iterative procedures must be used to obtain it. On the contrary, when we work with the π-parameterization, we can easily obtain the estimates of $\boldsymbol{\lambda}$ using (4) in a simple Monte Carlo scheme. In fact, for a given graph G we can always reconstruct the joint distribution $\boldsymbol{\pi}$ via $\boldsymbol{\pi}^G$ and then simply calculate the marginal log-linear parameters directly using (4).

4 Prior distributions on cell probabilities

In the following we work using conjugate priors on the probability parameters and then calculate the corresponding log-linear parameters using Monte Carlo schemes. In the section which follows we present prior set-ups based on Dirichlet distributions and ways to specify the prior parameters.

Table 1: Log-linear $\lambda-$parameterization for the graphs of Figure 1(a), (c), (d).

Independence model $\{A, B, C\}$	
Margins	Parameters
AB	$\lambda_{\emptyset}^{M_{AB}}$, $\lambda_{A}^{M_{AB}}$, $\lambda_{B}^{M_{AB}}$, $\lambda_{AB}^{M_{AB}} = 0$
BC	$\lambda_{C}^{M_{BC}}$, $\lambda_{BC}^{M_{BC}} = 0$
AC	$\lambda_{AC}^{M_{AC}} = 0$
ABC	$\lambda_{ABC}^{M_{ABC}} = 0$
Edge Model $\{AB, C\}$	
Margins	Parameters
AC	$\lambda_{\emptyset}^{M_{AC}}$, $\lambda_{A}^{M_{AC}}$, $\lambda_{C}^{M_{AC}}$, $\lambda_{AC}^{M_{AC}} = 0$
BC	$\lambda_{B}^{M_{BC}}$, $\lambda_{BC}^{M_{BC}} = 0$
ABC	$\lambda_{AB}^{M_{ABC}}$, $\lambda_{ABC}^{M_{ABC}} = 0$
Gamma Model $\{AB, BC\}$	
Margins	Parameters
AC	$\lambda_{\emptyset}^{M_{AC}}$, $\lambda_{A}^{M_{AC}}$, $\lambda_{C}^{M_{AC}}$, $\lambda_{AC}^{M_{AC}} = 0$
ABC	$\lambda_{B}^{M_{ABC}}$, $\lambda_{AB}^{M_{ABC}}$, $\lambda_{BC}^{M_{ABC}}$, $\lambda_{ABC}^{M_{ABC}}$

4.1 Conjugate Priors

For the specification of the prior distribution on the probability parameter vector we initially consider a Dirichlet distribution with parameters $\boldsymbol{\alpha} = \big(\alpha(i),\, i \in \mathcal{I}\,\big) = \big(\alpha_{abc},\, a = 1,\ldots,|\mathcal{I}_A|, b = 1,\ldots,|\mathcal{I}_B|, c = 1,\ldots,|\mathcal{I}_C|\,\big)$ for the vector of the joint probabilities of the full table $\boldsymbol{\pi}$. Hence, for the full table $\pi \sim \mathcal{D}i(\boldsymbol{\alpha})$ with prior density given by

$$\begin{aligned} f(\boldsymbol{\pi}) &= \frac{\Gamma(\alpha)}{\prod\limits_{i\in\mathcal{I}} \Gamma\Big(\alpha(i)\Big)} \prod_{i\in\mathcal{I}} \pi(i)^{\alpha(i)-1} \\ &= \frac{\Gamma(\alpha)}{\prod\limits_{a=1}^{|\mathcal{I}_A|}\prod\limits_{b=1}^{|\mathcal{I}_B|}\prod\limits_{c=1}^{|\mathcal{I}_C|} \Gamma\Big(\alpha_{abc}\Big)} \prod_{a=1}^{|\mathcal{I}_A|}\prod_{b=1}^{|\mathcal{I}_B|}\prod_{c=1}^{|\mathcal{I}_C|} \pi_{abc}^{\left(\alpha_{abc}\,-1\right)} = f_{\mathcal{D}i}(\boldsymbol{\pi};\, \boldsymbol{\alpha}) \end{aligned} \tag{5}$$

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

Under this set-up, the marginal prior of $\pi(i)$ is a Beta distribution with parameters $\alpha(i)$ and $\alpha - \alpha(i)$, i.e. $\pi(i) \sim Beta\Big(\alpha(i), \alpha - \alpha(i)\Big)$. The prior mean and variance of each cell is given by

$$E\big[\pi(i)\big] = \frac{\alpha(i)}{\alpha} \quad \text{and} \quad V\big[\pi(i)\big] = \frac{\alpha(i)\{\alpha - \alpha(i)\}}{\alpha^2(\alpha+1)}\ .$$

When no prior information is available then we usually set all $\alpha(i) = \frac{\alpha}{|\mathcal{I}|}$ resulting to

$$E\big[\pi(i)\big] = \frac{1}{|\mathcal{I}|} \quad \text{and} \quad V\big[\pi(i)\big] = \frac{|\mathcal{I}| - 1}{|\mathcal{I}|^2(\alpha+1)}\ .$$

Small values of α increase the variance of each cell probability parameter. Usual choices for α are the values $|\mathcal{I}|/2$ (Jeffrey's prior), $|\mathcal{I}|$ and 1 (corresponding to $\alpha(i)$ equal to 1/2, 1 and $1/|\mathcal{I}|$ respectively); for details see Dellaportas and Forster (1999). 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 (Lindley, 1957, Bartlett, 1957). Here this effect is not so adverse, as for example in usual variable selection for generalized linear models, for two reasons. Firstly even if we consider the limiting case where $\alpha(i) = \frac{\alpha}{|\mathcal{I}|}$ with $\alpha \to 0$, the variance is finite and equal to $(|\mathcal{I}| - 1)/|\mathcal{I}|^2$. Secondly, the distributions of all models are constructed from a common distribution of the full model/table making the prior distributions 'compatible' across different models (Dawid and Lauritzen, 2000 and Roverato and Consonni, 2004).

The model specific prior distributions are defined by the constraints imposed by the model's graphical structure and the adopted factorization. The prior distribution also factorizes in the same manner as the likelihood described in section 3.1. Thus, the prior for the saturated model is the usual Dirichlet given by (5).

Both the independence and the edge models can be expressed as product of independent Dirichlet distributions on probability parameters of the disconnected sets. Hence, for the independence model the prior is given by

$$f\left(\boldsymbol{\pi}^G|G\right) = \prod_{v\in\{A,B,C)} \left[\frac{\Gamma\left(\alpha_v\right)}{\prod\limits_{i_v\in\mathcal{I}_v} \Gamma\Big(\alpha_v(i_v)\Big)} \prod_{i_v\in\mathcal{I}_v} \pi_v(i_v)^{\alpha_v(i_v)-1} \right] = \prod_{v\in\{A,B,C)} f_{\mathcal{D}i}(\boldsymbol{\pi}_v;\ \boldsymbol{\alpha}_v)$$

while for the edge model of type $\{e, \overline{e}\}$ is given by

$$\begin{aligned} f\left(\boldsymbol{\pi}^G|G\right) &= \frac{\Gamma\left(\alpha_e\right)}{\prod\limits_{i_e\in\mathcal{I}_e} \Gamma\Big(\alpha_e(i_e)\Big)} \prod_{i_e\in\mathcal{I}_e} \pi_e(i_e)^{\alpha_e(i_e)-1} \times \frac{\Gamma\left(\alpha_{\overline{e}}\right)}{\prod\limits_{i_{\overline{e}}\in\mathcal{I}_{\overline{e}}} \Gamma\Big(\alpha_{\overline{e}}(i_{\overline{e}})\Big)} \prod_{i_{\overline{e}}\in\mathcal{I}_{\overline{e}}} \pi_{\overline{e}}(i_{\overline{e}})^{\alpha_{\overline{e}}(i_{\overline{e}})-1} \\ &= f_{\mathcal{D}i}(\boldsymbol{\pi}_e;\ \boldsymbol{\alpha}_e) \times f_{\mathcal{D}i}(\boldsymbol{\pi}_{\overline{e}};\ \boldsymbol{\alpha}_{\overline{e}}) \end{aligned}$$

with $\alpha_M = \sum\limits_{i_M\in\mathcal{I}_M} \alpha_M(i_M)$ and $\boldsymbol{\alpha}_M = \Big(\alpha_M(i_M); M \in \mathcal{I}_M\Big)$ for any $M \subseteq \mathcal{V}$. For example, for the edge model $\{AB, C\}$ with $e = C$ the prior will be a product of a Dirichlet distributions for $\boldsymbol{\pi}_{AB}$ and $\boldsymbol{\pi}_C$ with parameters $\boldsymbol{\alpha}_{AB}$ and $\boldsymbol{\alpha}_C$ respectively.

The prior can be written as

$$f\left(\boldsymbol{\pi}^G|G\right) = \frac{\Gamma\left(\sum_{c=1}^{|\mathcal{I}_C|}\alpha_C(c)\right)}{\prod_{c=1}^{|\mathcal{I}_C|}\Gamma\Big(\alpha_C(c)\Big)}\prod_{c=1}^{|\mathcal{I}_C|}\pi_{++c}^{\left(\alpha_C(c)-1\right)} \times \frac{\Gamma\left(\sum_{a=1}^{|\mathcal{I}_A|}\sum_{b=1}^{|\mathcal{I}_B|}\alpha_{AB}(ab)\right)}{\prod_{a=1}^{|\mathcal{I}_A|}\prod_{b=1}^{|\mathcal{I}_B|}\Gamma\Big(\alpha_{AB}(ab)\Big)}\prod_{a=1}^{|\mathcal{I}_A|}\prod_{b=1}^{|\mathcal{I}_B|}\pi_{ab+}^{\left(\alpha_{AB}(ab)-1\right)}.$$

Since the above prior densities are a product of Dirichlet distributions (over the parameters of all disconnected sets) we denote them by

$$f\left(\boldsymbol{\pi}^G\right) = f_{\mathcal{PD}}\Big(\boldsymbol{\pi}_d;\ \boldsymbol{\alpha}_d,\, d\in\mathcal{D}(G)\Big). \tag{6}$$

For the gamma structure the prior is given by

$$p\left(\boldsymbol{\pi}^G\right) = \left\{\prod_{i_{\overline{v}_c}\in\mathcal{I}_{\overline{v}_c}} f_{\mathcal{D}i}\Big(\boldsymbol{\pi}_{v_c|\overline{v}_c}(\cdot|i_{\overline{v}_c});\ \boldsymbol{\alpha}_{v_c|\overline{v}_c}(\cdot|i_{\overline{v}_c})\Big)\right\} f_{\mathcal{D}i}\Big(\boldsymbol{\pi}_{v_1};\ \boldsymbol{\alpha}_{v_1}\Big) f_{\mathcal{D}i}\Big(\boldsymbol{\pi}_{v_2};\ \boldsymbol{\alpha}_{v_2}\Big). \tag{7}$$

with $\boldsymbol{\pi}_{v_c|\overline{v}_c}(\cdot|i_{\overline{v}_c}) = \Big(\boldsymbol{\pi}_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c}), i_{v_c}\in\mathcal{I}_{v_c}\Big)$ and $\boldsymbol{\alpha}_{v_c|\overline{v}_c}(\cdot|i_{\overline{v}_c}) = \Big(\boldsymbol{\alpha}_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c}), i_{v_c}\in\mathcal{I}_{v_c}\Big)$. The fist part of equation (7), that is the product for all level of $\overline{v}_c$ of Dirichlet distributions of the conditional probabilities, can be denoted by $f_{\mathcal{CPD}}\big(\boldsymbol{\pi}_{v_c|\overline{v}_c};\ \boldsymbol{\alpha}\big)$. Then, the prior density (7) can be written as

$$f\left(\boldsymbol{\pi}^G\right) = f_{\mathcal{CPD}}\big(\boldsymbol{\pi}_{v_c|\overline{v}_c};\ \boldsymbol{\alpha}_{v_c|\overline{v}_c}\big) f_{\mathcal{PD}}\Big(\boldsymbol{\pi}_v;\ \boldsymbol{\alpha}_v,\, v\in\overline{v}_c\Big), \tag{8}$$

where $\boldsymbol{\alpha}_{v_c|\overline{v}_c} = \Big(\boldsymbol{\alpha}_{v_c|\overline{v}_c}(i_{v_c}|i_{\overline{v}_c}), i_{v_c}\in\mathcal{I}_{v_c}, i_{\overline{v}_c}\in\mathcal{I}_{\overline{v}_c}\Big)$. For example if $v_c = B$ then the prior can be written as

$$\begin{aligned} p\left(\boldsymbol{\pi}^G\right) &= \prod_{a=1}^{|\mathcal{I}_A|}\prod_{c=1}^{|\mathcal{I}_C|}\left\{\frac{\Gamma\Big(\alpha_{B|AC=ac}\Big)}{\prod_{b=1}^{|\mathcal{I}_B|}\Gamma\Big(\alpha_{B|AC}(b|ac)\Big)}\left(\prod_{b=1}^{|\mathcal{I}_B|}\left[\pi_{B|AC}(b|ac)\right]^{\alpha_{B|AC}(b|ac)-1}\right)\right\} \\ &\times \frac{\Gamma\left(\alpha_A\right)}{\prod_{a=1}^{|\mathcal{I}_A|}\Gamma\Big(\alpha_A(a)\Big)}\prod_{a=1}^{|\mathcal{I}_A|}\pi_{a++}^{\left(\alpha_A(a)-1\right)} \times \frac{\Gamma\left(\alpha_C\right)}{\prod_{c=1}^{|\mathcal{I}_C|}\Gamma\Big(\alpha_C(c)\Big)}\prod_{c=1}^{|\mathcal{I}_C|}\pi_{++c}^{\left(\alpha_C(c)-1\right)} \end{aligned}$$

where $\alpha_{B|AC=ac} = \sum_{b=1}^{|\mathcal{I}_B|}\alpha_{B|AC}(b|ac)$.

4.2 Compatible Prior Distributions

In order to make the prior distributions 'compatible' across models, we define the prior parameters of $\boldsymbol{\pi}^G$ from the corresponding parameters of the prior distribution (5) imposed on the probabilities $\boldsymbol{\pi}$ of the full table; see Dawid and Lauritzen (2000), Roverato and Consonni (2004).

Let us consider a marginal $M \in \mathcal{M}(G)$ for which we wish to estimate the probability parameters $\boldsymbol{\pi}_M = (\pi_M(i_M), i_M \in \mathcal{I}_M)$. The resulting prior is $\boldsymbol{\pi}_M \sim \mathcal{D}i\big(\boldsymbol{\alpha}_M\big)$, that is a Dirichlet distribution with parameters $\boldsymbol{\alpha}_M = (\alpha_M(i_M), i_M \in \mathcal{I}_M)$ given by

$$\alpha_M(i_M) = \sum_{j \in \mathcal{I}_{\overline{M}}} \alpha(i_M, j_{\overline{M}}) = \sum_{\{j \in \mathcal{I}: j_M = i_M\}} \alpha(j),$$

see (i) of Lemma 7.2 in Dawid and Lauritzen (1993, p.1304).

For example, consider a three way table with $\mathcal{V} = \{A, B, C\}$ and the marginal $M = C$. Then the prior imposed on the parameters $\boldsymbol{\pi}_C$ of the marginal C is given by

$$\boldsymbol{\pi}_C \sim \mathcal{D}i\big(\boldsymbol{\alpha}_C\big) \quad \text{with} \quad \alpha_c = \alpha_{++c} = \sum_{a=1}^{|\mathcal{I}_A|} \sum_{b=1}^{|\mathcal{I}_B|} \alpha_{abc} \text{ for } c = 1, 2, \ldots, |\mathcal{I}_C|,$$

where $\boldsymbol{\alpha}_C = \Big(\alpha_{++c}, c = 1, \ldots, |\mathcal{I}_C|\Big)$. Also note that under this prior set-up

$$\alpha_M \;\; = \sum_{i_M \in \mathcal{I}_M} \alpha_M(i_M) = \sum_{i_M \in \mathcal{I}_M} \sum_{j_M \in \mathcal{I}_{\overline{M}}} \alpha(i_M, j_M)$$

For the conditional distribution of $M_1|M_2$ with $M_1 \neq M_2 \in \mathcal{M}(G)$ we work in a similar way. The vector $\boldsymbol{\pi}_{M_1|M_2}(\cdot|i_{M_2}) = \big(\pi_{M_1|M_2}(i_{M_1}|i_{M_2}), i_{M_1} \in \mathcal{I}_{M_1}\big)$ a priori follows a Dirichlet distribution

$$\boldsymbol{\pi}_{M_1|M_2}(\cdot|i_{M_2}) \sim \mathcal{D}i\Big(\alpha_{M_1 \cup M_2}(i_{M_1}, i_{M_2}),\, i_{M_1} \in \mathcal{I}_{M_1}\Big).$$

The above structure derives from the decomposition of a Dirichlet as a ratio of Gamma distributions; see also Lemma 7.2 (ii) in Dawid and Lauritzen (1993, p.1304).

For example, consider marginals $M_1 = A$ and $M_2 = B$ in a three way contingency table with $\mathcal{V} = \{A, B, C\}$. Then, for a specific level of variable B, say $i_B = 2$,

$$\boldsymbol{\pi}_{A|B}(\cdot|i_B = 2) \sim \mathcal{D}i\big(\boldsymbol{\alpha}_{AB}(\cdot, 2)\big)$$

where $\boldsymbol{\alpha}_{AB}(\cdot, 2) = \Big(\alpha_{a2+}, a = 1, \ldots, |\mathcal{I}_A|\Big)$ and $\alpha_{a2+} = \sum_{c=1}^{|\mathcal{I}_C|} \alpha_{a2c}$.

4.3 Specification of Prior Parameters Using Imaginary Data.

In graphical model literature there is a debate about the use of conjugate priors based on Dirichlet distributions; see for example in Steck and Jaakkola (2002), Steck (2008) and Ueno (2008). As pointed out in Section 4.1, the parameters of the Dirichlet prior should be carefully specified. In order to do this we adopt ideas based on the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000). We use their approach to advocate sensible values for the Dirichlet prior parameters on the full table and the corresponding induced values for the rest of the graphs as described in the previous sub-section. Although here we restrict our attention to marginal independence graphs, the procedure can be applied also for prior elicitation for undirected graphs and DAGs.

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

$$\begin{aligned} f(\boldsymbol{\pi}) &\propto f(\boldsymbol{n}^*|\boldsymbol{\pi})^w \times f_{\mathcal{D}i}\big(\boldsymbol{\pi};\ \alpha(i) = \alpha_0, i \in \mathcal{I}\big) \\ &\propto \prod_{i \in \mathcal{I}} \pi(i)^{w\, n^*(i) + \alpha_0 - 1} \\ &= f_{\mathcal{D}i}\big(\boldsymbol{\pi};\ \alpha(i) = w\, n^*(i) + \alpha_0, i \in \mathcal{I}\big)\ . \end{aligned} \tag{9}$$

Using the above prior set up, we expect a priori to observe a total number of $w\,N^* + |\mathcal{I}|\alpha_0$ observations. The parameter w is used to specify the steepness of the prior distribution and the weight of belief on each prior observation. For $w = 1$ then each imaginary observation has the same weight as the actual observations. Values of $w < 1$ will give less weight to each imaginary observation while $w > 1$ will increase the weight of believe on the prior/imaginary data. Overall the prior will account for the $(w\,N^* + |\mathcal{I}|\alpha_0)/(w\,N^* + N + |\mathcal{I}|\alpha_0)$ of the total information used in the posterior distribution. Hence for $w = 1$, $N^* = N$ and $\alpha_0 \to 0$ then both the prior and data will account for 50% of the information used in the posterior.

For $w = 1/N^*$ then $\alpha(i) = p^*(i) + \alpha_0$ with $p^*(i) = n^*(i)/N^*$, the prior data $\boldsymbol{n}^*$ will account for information of one data point while the total weight of the prior will be equal to $(1+|\mathcal{I}|\alpha_0)/(1+N+|\mathcal{I}|\alpha_0)$. If we further set $\alpha_0 = 0$, then the prior distribution (9) will account for information equivalent to a single observation. This prior set-up will be referred in this paper as the unit information prior (UIP). When no information is available, then we may further consider the choice of equal cell frequencies $n^*(i) = n^*$ for the imaginary data in order to support the simplest possible model under consideration. Under this approach $N^* = n^* \times |\mathcal{I}|$ and $w = 1/N^* = \frac{1}{n^* \times |\mathcal{I}|}$ resulting to

$$\boldsymbol{\pi} \sim \mathcal{D}i\Big(\alpha(i) = 1/|\mathcal{I}|,\, i \in \mathcal{I}\Big)\ .$$

The latter prior is equivalent to the one advocated by Perks (1947). It has the nice property that the prior on the marginal parameters does not depend on the size of the table; for example, for a binary variable, this prior will assign a $Beta(1/2,\ 1/2)$ prior on the corresponding marginal regardless the size of the table we work with (for example if we work with 2^3 or $2 \times 4 \times 5 \times 4$ table). This property is retained for any

prior distribution of type (9) with $w^* = 1/N^*$, $p^*(i) = 1/|\mathcal{I}|$ and $\alpha_0 \propto 1/|\mathcal{I}|$.

4.4 Comparison of Prior Set-ups

Since Perks' prior (with $\alpha(i) = 1/|\mathcal{I}|$) has a unit information interpretation, it can be used as a yardstick in order to identify and interpret the effect of any other prior distribution used. Prior distributions with $\alpha(i) < 1/|\mathcal{I}|$, or equivalently $\alpha < 1$, result in larger variance than the one imposed by our proposed unit information prior and hence they a posteriori supports more parsimonious models. On the contrary, prior distributions with $\alpha(i) > 1/|\mathcal{I}|$, or $\alpha > 1$, result in lower prior variance and hence they a posteriori support models with more complicated graph structure. So the variance ratio between a Dirichlet prior with $\alpha(i) = \alpha/|\mathcal{I}|$ and Perks prior is equal to

$$VR = \frac{V\left(\pi(i)\middle|\ \alpha(i) = \frac{\alpha}{|\mathcal{I}|}\right)}{V\left(\pi(i)\middle|\ \alpha(i) = |\mathcal{I}|^{-1}\right)} = \frac{2}{\alpha + 1}\ .$$

Table 2 presents the comparison of the information from the following prior choices:

(i) the Jeffrey's prior with $\alpha(i) = 1/2$;

(ii) the Unit Expected Cells prior (UEC) with $\alpha(i) = 1$;

(iii) the Unit Information Prior (UIP) which is derived by a power prior with $\alpha(i) = p^*(i)$, $w = 1/N^*$ and $a_0 = 0$; where $p^*(i)$ is the sample proportion of cell i estimated from a set of imaginary data $n^*(i)$;

  (a) Perks' prior (UIP-Perks') with $\alpha(i) = 1/|\mathcal{I}|$ which is equivalent to UIP coming from a table of imaginary data with all cell frequencies equal to one;

  (b) the Unit Information Empirical Bayes Prior (UI-EBP), which is derived by UIP with $p^*(i)$ set equal to the sample proportions $p(i) = n(i)/N$.

Table 2: Table of Prior Variance in Comparison to the Unit Information Prior (last row of the table)

			Variance Ratio		
	Parameter	$V\big[\pi(i)\big]$	General Equation	$2 \times 2 \times 2$	$3 \times 2 \times 4$
Jeffrey's	$\alpha(i) = 1/2$	$2\ \frac{\lvert\mathcal{I}\rvert - 1}{\lvert\mathcal{I}\rvert^2\{\lvert\mathcal{I}\rvert + 2\}}$	$4/\big(\lvert\mathcal{I}\rvert + 2\big)$	0.4	0.15
Unit Expected Cell (UEC)	$\alpha(i) = 1$	$\frac{\lvert\mathcal{I}\rvert - 1}{\lvert\mathcal{I}\rvert^2\{\lvert\mathcal{I}\rvert + 1\}}$	$2/\big(\lvert\mathcal{I}\rvert + 1\big)$	0.22	0.08
Unit Information Prior (UIP)	$\alpha(i) = p^*(i)$	$\frac{1}{2}p^*(i)\Big(1 - p^*(i)\Big)$	$\frac{\lvert\mathcal{I}\rvert^2}{\lvert\mathcal{I}\rvert - 1}p^*(i)\Big(1 - p^*(i)\Big)$	$9.14 \times p^*(i)\Big(1 - p^*(i)\Big)$	$25.04 \times p^*(i)\Big(1 - p^*(i)\Big)$
Perks' Prior (UIP-Perks')	$\alpha(i) = 1/\lvert\mathcal{I}\rvert$	$\frac{\lvert\mathcal{I}\rvert - 1}{2\lvert\mathcal{I}\rvert^2}$	1	1	1
Unit Information Empirical Bayes Prior (UI-EBP)	$\alpha(i) = p(i)$	$\frac{1}{2}p(i)\Big(1 - p(i)\Big)$	$\frac{\lvert\mathcal{I}\rvert^2}{\lvert\mathcal{I}\rvert - 1}V[\pi(i)]$	$9.14 \times p(i)\Big(1 - p(i)\Big)\Big)$	$25.04 \times p(i)\Big(1 - p(i)\Big)\Big)$

From this table, we observe that Jeffreys' prior variance is lower than the corresponding Perks' prior reaching a reduction of about 60% and 85% for a 2^3 and a $2 \times 3 \times 4$ table respectively. The reduction is even greater for the Unit Expected Cell prior reaching 78% and 92% respectively.

Finally, for the Empirical Bayes prior, based on the UIP approach, the prior variance for each $\pi(i)$ is equal to $V[\pi(i)] = \frac{1}{2}p(i)\big(1 - p(i)\big)$. Hence it depends on the observed proportion and can vary from zero (if $p(i) = 0$ or 1) to 1/8 if $p(i) = 1/2$. For values in the interval $(0.058, 0.942)$ the variance of the UI-EBP is higher than the corresponding UIP variance reaching its maximum when $p(i) = 1/2$ where it is 4.6 times the corresponding UIP prior variance. For $p(i) = 0.058$ or 0.942 the variances of the UIP and UI-EBP are equal while for the remaining values, UIP variance is higher.

5 Posterior Distributions

5.1 Posterior Distributions of Model Parameters

Since the prior is conjugate to the likelihood, the posterior can be derived easily as follows. For the saturated model the posterior distribution is also a Dirichlet distribution

$$f(\boldsymbol{\pi}\big|\boldsymbol{n}, G_S) = f_{\mathcal{D}i}\big(\boldsymbol{\pi}; \widetilde{\boldsymbol{\alpha}}\big)$$

with parameters

$$\widetilde{\boldsymbol{\alpha}} = \big(\widetilde{\alpha}(i) = \alpha(i) + n(i),\ i \in \mathcal{I}\big) = \big(\widetilde{\alpha}_{abc} = \alpha_{abc} + n_{abc}, a \in \mathcal{I}_A, b \in \mathcal{I}_B, c \in \mathcal{I}_C\big).$$

For the independence and the edge structure the density of the posterior distribution is is equivalent to (6),

$$f(\boldsymbol{\pi}^G\big|\boldsymbol{n}, G) = f_{\mathcal{PD}}\big(\boldsymbol{\pi}^G; \widetilde{\boldsymbol{\alpha}}^G\big)$$

with

$$\widetilde{\boldsymbol{\alpha}}^G = \Big(\widetilde{\boldsymbol{\alpha}}_d,\, d \in \mathcal{D}(G)\Big) \;\text{ and }\; \widetilde{\boldsymbol{\alpha}}_d = \Big(\widetilde{\alpha}_d(i_d) = \alpha_d(i_d) + n_d(i_d),\; i_d \in \mathcal{I}_d\Big).$$

Hence for the independence model the parameter vector of the posterior distribution is given by $\widetilde{\boldsymbol{\alpha}}^G = \Big(\widetilde{\boldsymbol{\alpha}}_A, \widetilde{\boldsymbol{\alpha}}_B, \widetilde{\boldsymbol{\alpha}}_C\Big)$ with $\widetilde{\boldsymbol{\alpha}}_A = \Big(\widetilde{\alpha}_{a++} = \alpha_{a++} + n_{a++},\, a = 1, \ldots, |\mathcal{I}_A|\Big)$, $\widetilde{\boldsymbol{\alpha}}_B = \Big(\widetilde{\alpha}_{+b+} = \alpha_{+b+} + n_{+b+},\, b = 1, \ldots, |\mathcal{I}_B|\Big)$ and $\widetilde{\boldsymbol{\alpha}}_C = \Big(\widetilde{\alpha}_{++c} = \alpha_{++c} + n_{++c},\, c = 1, \ldots, |\mathcal{I}_C|\Big)$ while for the edge model $\{AB, C\}$ with $e = C$ the parameters vector is given by $\widetilde{\boldsymbol{\alpha}}^G = \Big(\widetilde{\boldsymbol{\alpha}}_{AB}, \widetilde{\boldsymbol{\alpha}}_C\Big)$ with $\widetilde{\boldsymbol{\alpha}}_{AB} = \Big(\widetilde{\alpha}_{ab+} = \alpha_{ab+} + n_{ab+},\, a = 1, \ldots, |\mathcal{I}_A|,\, b = 1, \ldots, |\mathcal{I}_B|\Big)$ and $\widetilde{\boldsymbol{\alpha}}_C$ as above.

Finally, for the gamma structure

$$f(\boldsymbol{\pi}^G | \boldsymbol{n}, G) = f_{\mathcal{CPD}}\big(\boldsymbol{\pi}_{v_c|v_{\overline{c}}};\, \widetilde{\boldsymbol{\alpha}}\big) \times f_{\mathcal{PD}}\big(\boldsymbol{\pi}_v;\, \widetilde{\boldsymbol{\alpha}}_v,\, v \in \overline{v}_c\big)$$

i.e. a distribution with density equivalent to the corresponding prior (7) with parameters $\widetilde{\boldsymbol{\alpha}}^G = \big(\widetilde{\boldsymbol{\alpha}},\, \widetilde{\boldsymbol{\alpha}}_v,\, v \in \overline{v}_c\big)$.

Finally, from the properties of the Dirichlet distribution we can derive the marginal posterior distribution of each element of $\boldsymbol{\pi}^G$ which is a Beta distribution with the appropriate parameters (see Section 4.1).

5.2 Marginal Likelihood of Each Graph

For model choice we need to estimate the posterior model probabilities $f(G|n) \propto f(\boldsymbol{n}|G)f(G)$, with $f(\boldsymbol{n}|G)$ marginal likelihood of the model and $f(G)$ prior distribution on G. Here we restrict to the simple case where $f(G)$ is uniform, hence the posterior will depend only on the marginal likelihood $f(\boldsymbol{n}|G)$ of the model under consideration. The marginal likelihood can be calculated analytically since the above prior set-up is conjugate.

For the saturated model the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = \mathsf{C}(\boldsymbol{n}) \times \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}})}{\mathsf{B}(\boldsymbol{\alpha})}$$

where $\mathsf{C}(\boldsymbol{n})$ is the usual multinomial constant and $\mathsf{B}(\boldsymbol{\alpha})$ is the normalizing constant of the multinomial beta function given by

$$\mathsf{B}\big(\boldsymbol{\alpha}\big) = \frac{\prod\limits_{i\in\mathcal{I}} \Gamma\Big(\alpha(i)\Big)}{\Gamma\left(\sum\limits_{i\in\mathcal{I}} \alpha(i)\right)} \ .$$

respectively.

For the independence model the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = \mathsf{C}(\boldsymbol{n}) \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_A)}{\mathsf{B}(\boldsymbol{\alpha}_A)} \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_B)}{\mathsf{B}(\boldsymbol{\alpha}_B)} \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_C)}{\mathsf{B}(\boldsymbol{\alpha}_C)},$$

while for the edge model $\{e, \overline{e}\}$ the marginal likelihood is calculated by

$$f(\boldsymbol{n}|G) = \mathsf{C}(\boldsymbol{n}) \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_e)}{\mathsf{B}(\boldsymbol{\alpha}_e)} \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_{\overline{e}})}{\mathsf{B}(\boldsymbol{\alpha}_{\overline{e}})} \ .$$

For example, for $e = C$ the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = \mathsf{C}(\boldsymbol{n}) \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_C)}{\mathsf{B}(\boldsymbol{\alpha}_C)} \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_{AB})}{\mathsf{B}(\boldsymbol{\alpha}_{AB})} \ .$$

Finally, for the gamma structure the marginal likelihood $f(\boldsymbol{n}|G)$ is given by

$$f(\boldsymbol{n}|G) \quad = \quad \mathsf{C}(\boldsymbol{n}) \prod_{i_{\overline{v}_c} \in \mathcal{I}_{\overline{v}_c}} \frac{\mathsf{B}\big(\widetilde{\boldsymbol{\alpha}}(\cdot, i_{\overline{v}_c})\big)}{\mathsf{B}\big(\boldsymbol{\alpha}(\cdot, i_{\overline{v}_c})\big)} \prod_{v \in \overline{v}_c} \frac{\mathsf{B}(\widetilde{\boldsymbol{\alpha}}_v)}{\mathsf{B}(\boldsymbol{\alpha}_v)} \ . \tag{10}$$

5.3 Estimation of the Posterior Distribution of Marginal Association Log-Linear Parameters

The posterior distribution of the marginal log-linear parameters $\boldsymbol{\lambda}^G$ can be estimated in a straightforward manner using Monte Carlo samples from the posterior distribution of $\boldsymbol{\pi}^G$. Specifically, a sample from the posterior distribution of $\boldsymbol{\lambda}^G$ can be generated by the following steps.

i) Generate a random sample $\boldsymbol{\pi}^{G,(t)}(t = 1, \ldots, T)$ from the posterior distribution of $\boldsymbol{\pi}^G$.

ii) At each iteration t, calculate the full table of probabilities $\boldsymbol{\pi}^{(t)}$ from $\boldsymbol{\pi}^{G,(t)}$.

iii) The vector of marginal log-linear parameters, $\boldsymbol{\lambda}^{G,(t)}$, can be easily obtained from $\boldsymbol{\pi}^{(t)}$ via equation (4) which becomes

$$\boldsymbol{\lambda}^{G,(t)} = \boldsymbol{C}^G \log\left(\boldsymbol{M}^G \mathrm{vec}\left(\boldsymbol{\pi}^{(t)}\right)\right)$$

where $\boldsymbol{C}^G$ and $\boldsymbol{M}^G$ are the contrast and marginalization matrices under graph G. Note that some elements of $\boldsymbol{\lambda}^G$ will automatically be constrained to zero for all generated values due to the graphical structure of the model G and the way we calculate log-linear parameters using the previous equation.

Finally, we can use the generated values $\left(\boldsymbol{\lambda}^{G,(t)}; t = 1, 2, \ldots, T\right)$ to estimate summaries of the posterior distribution $f(\boldsymbol{\lambda}^G|G)$ or obtain plots fully describing this distribution.

6 Illustrative examples

The methodology described in the previous sections is now illustrated on two real data sets, a $2 \times 2 \times 2$ and a $3 \times 2 \times 4$ tables. In both example we compare the results obtained with our yardstick prior, the UIP-Perks' prior ($\alpha(i) = 1/|I|$), with those obtained using Jeffrey's ($\alpha(i) = 1/2$), Unit Expected Cell ($\alpha(i) = 1$), and unit information Empirical Bayes ($\alpha(i) = p(i)$) priors.

6.1 A $2 \times 2 \times 2$ Table: Antitoxin Medication Data

We consider a data set presented by Healy (1988) regarding a study on the relationship between patient condition (more or less severe), assumption of antitoxin (yes or not) and survival status (survived or not); see Table 3. In Table 4 we compare posterior model probabilities under the four different prior set-ups.

Table 3: Antitoxin data

		Survival (S)	
Condition (C)	Antitoxin (A)	No	Yes
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Under all prior assumptions the maximum a posteriori model (MAP) is SC+A (we omit the conventional crossing (*) operator between variables for simplicity), assuming the marginal independence of Antitoxin from the remaining two variables.

Under Empirical Bayes and UIP-Perks' priors the posterior distribution is concentrate on the MAP model (it takes into account 93.4% and 91.7% respectively of the posterior model probabilities). The posterior distributions under the Jeffreys' and the unit expected prior set-ups are more disperse, supporting the three models (SC+A, AS+SC and ASC) with posterior weights higher than 10% and accounting around the 94% of the posterior model probabilities. Model $AS+SC$ is also the model with the second highest posterior probability under UIP-Perks' prior but its weight is considerably

lower than the corresponding probability of the MAP model.

Table 4: Posterior model probabilities (%) for the Antitoxin data.

		Model							
		A+S+C	AS+C	AC+S	SC+A	AS+AC	AS+SC	AC+SC	ASC
Prior Distribution	α_{abc}	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Jeffreys'	1/2	0.3	1.5	0.2	59.7	0.1	21.7	3.0	13.4
Unit Expected Cell	1	0.2	1.1	0.2	37.2	0.1	30.2	4.7	26.2
Empirical Bayes	$p(i)$	1.6	2.4	0.3	93.4	0.0	1.7	0.2	0.4
UIP-Perks'	$1/\|I\|$	1.2	2.1	0.3	91.7	0.0	3.5	0.4	0.8

Figure 2 presents boxplots summarizing 2.5%, 97.5% posterior percentiles and quantiles of the joint probabilities for the MAP model (SC+A) for the four prior set-ups. Since direct calculation from the posterior distribution is not feasible, we estimated the posterior summaries via Monte Carlo simulation (1000 values). From this figure, we observe minor differences between the posterior distributions obtained under the UIP-Perks' and the empirical Bayes prior. More differences are observed between Perks' UIP and the posterior distributions under the two other prior set-ups. Differences are higher for the first two cell probabilities, i.e. for $\pi(1,1,1)$ and $\pi(2,1,1)$.

Similarly in Figure 3 we present boxplots providing posterior summaries for models $SC + A$, $AS + SC$ and ASC under the UIP-Perks' prior set-up. The first two models are the ones with highest posterior probabilities and all of their summaries have been calculated using Monte Carlo simulation (1000 values). The saturated was used mainly as reference model since it is the only one for which the posterior distributions are available analytically. From the figure we observe that the posterior distributions on the joint probabilities $\boldsymbol{\pi}$ of the full table are quite different highly depending on the assumed model structure.

Figure 2: Antitoxin data: Boxplots summarizing 2.5%, 97.5% posterior percentiles and quantiles of the joint probabilities $\pi_{ABC}(i, j, k)$ for the MAP model (SC+A) for all prior set-ups (J=Jeffreys', U=Unit Expected Cell, E=Empirical Bayes, P=Perks') .

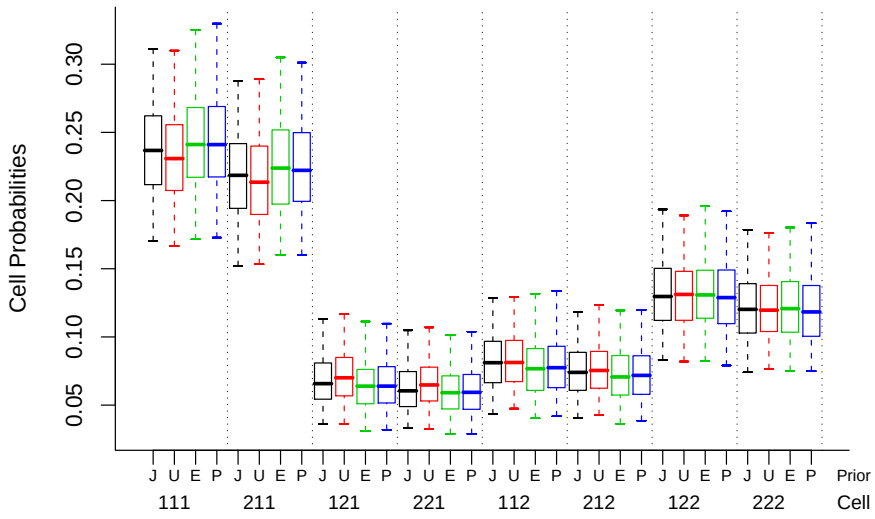
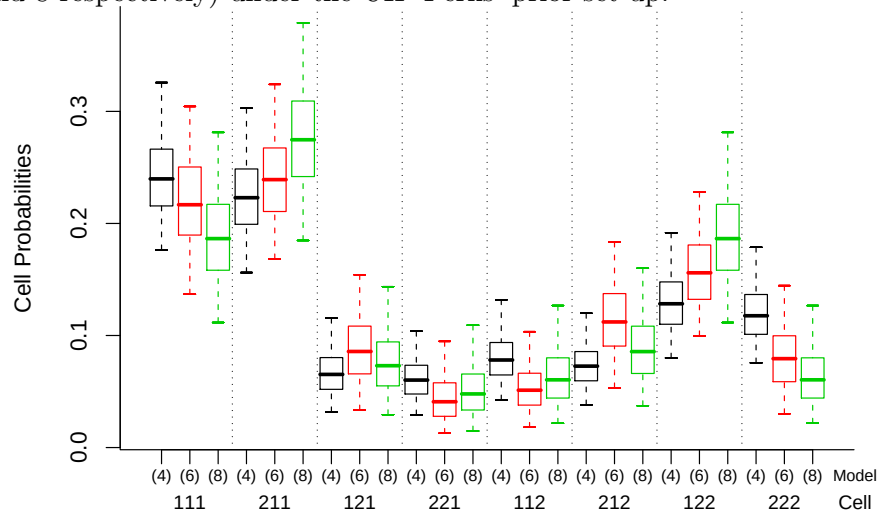


Figure 3: Antitoxin data: Boxplots summarizing 2.5%, 97.5% posterior percentiles and quantiles of the joint probabilities $\pi_{ABC}(i, j, k)$ for models SC+A, AS+SC and ASC (4, 6 and 8 respectively) under the UIP-Perks' prior set-up.



Finally, posterior summaries for the probability parameters $\boldsymbol{\pi}^G$ and the marginal log-linear parameters $\boldsymbol{\lambda}^G$ for models $SC + A$, $AS + SC$ and ASC (as described above) under the UIP-Perks' prior are provided in Tables 5 and 6 respectively. All summaries of each element of $\boldsymbol{\pi}^G$ are obtained analytically based on the Beta distribution induced by the corresponding Dirichlet posterior distributions of $\boldsymbol{\pi}^G$. Posterior summaries of $\boldsymbol{\lambda}^G$ are estimated using the Monte Carlo strategy (1000 values) discussed in section 5. As commented in this section, some elements of $\boldsymbol{\lambda}^G$ for graphs $SC+A$ and $AS+SC$ are constrained to zero due the way we have constructed our model. Hence for $SC+A$, the maximal interaction terms for the disconnected sets AS, AC and ASC, i.e. parameters $\lambda_{AS}(2,2)$, $\lambda_{AC}(2,2)$ and $\lambda_{ASC}(2,2,2)$, are constrained to be zero for all generated observations. Similar is the picture for model $AS + SC$, but now only marginals AC and ABC correspond to disconnected sets implying that $\lambda_{AC}(2,2) = \lambda_{ASC}(2,2,2) = 0$.

6.2 A $3 \times 2 \times 4$ table: Alcohol Data

Here we analyze a well known data set presented by Knuiman and Speed (1988) regarding a small study held in Western Australia on the relationship between Alcohol intake (A), Obesity (O) and High blood pressure (H); see Table 7.

In Table 8 we report posterior model probabilities and corresponding log-marginal likelihoods for each models. Under all prior set-ups the posterior model probability is concentrated on models H+A+O, HA+O and HO+A. Empirical Bayes and UIP-Perks' support the independence model (with posterior model probability of 0.878 and 0.807 respectively) whereas Jeffreys' and Unit Expected support a more complex structure, HO+A (with posterior model probability of 0.837 and 0.859 respectively).

Table 5: Antitoxin data: Posterior summaries of model parameters for models $SC + A$, $AS + SC$ and ASC under the UIP-Perks' prior set-up; $\widetilde{a}$ and $\widetilde{b}$ are the parameters of the resulted Beta marginal posterior distribution.

Model 4: $SC + A$					Beta Posterior Parameters	
Parameter	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$	$\widetilde{a}$	$\widetilde{b}$
$\pi_{SC}(1,1)$	0.47	0.055	0.36	0.57	37.25	42.75
$\pi_{SC}(2,1)$	0.13	0.037	0.06	0.21	10.25	69.75
$\pi_{SC}(1,2)$	0.15	0.040	0.08	0.24	12.25	67.75
$\pi_{A}(1)$	0.52	0.056	0.41	0.63	41.50	38.50
Model 6: $AS + SC$						
Parameter	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$	$\widetilde{a}$	$\widetilde{b}$
$\pi_{S\|AC}(1\|1,1)$	0.71	0.096	0.51	0.88	15.12	6.12
$\pi_{S\|AC}(1\|2,1)$	0.84	0.070	0.68	0.95	22.12	4.12
$\pi_{S\|AC}(1\|1,2)$	0.25	0.094	0.09	0.46	5.12	15.12
$\pi_{S\|AC}(1\|2,2)$	0.58	0.136	0.31	0.83	7.12	5.12
$\pi_{A}(1)$	0.52	0.056	0.41	0.63	41.50	38.50
$\pi_{C}(1)$	0.59	0.055	0.48	0.70	47.50	32.50
Model 8: ASC (Saturated)						
Parameter	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$	$\widetilde{a}$	$\widetilde{b}$
$\pi(1,1,1)$	0.19	0.044	0.11	0.28	15.12	64.88
$\pi(2,1,1)$	0.28	0.050	0.18	0.38	22.12	57.88
$\pi(1,2,1)$	0.08	0.030	0.03	0.14	6.12	73.88
$\pi(2,2,1)$	0.05	0.025	0.01	0.11	4.12	75.88
$\pi(1,1,2)$	0.06	0.027	0.02	0.13	5.12	74.88
$\pi(2,1,2)$	0.09	0.032	0.04	0.16	7.12	72.88
$\pi(1,2,2)$	0.19	0.044	0.11	0.28	15.12	64.88
$\pi(2,2,2)$	0.06	0.027	0.02	0.13	5.12	74.88

Table 6: Antitoxin data: Posterior summaries for lambda for models SC+A, AS+SC and ASC under the UIP-Perks' prior set-up

Model 4: $SC + A$					
Parameter	Marginal table	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$
$\lambda_{\emptyset}$	M_{AS}	-1.429	0.032	-1.513	-1.388
$\lambda_A(2)$	M_{AS}	-0.040	0.113	-0.258	0.181
$\lambda_S(2)$	M_{AS}	-0.245	0.118	-0.480	-0.021
$\lambda_{AS}(2,2)$	M_{AS}	0.000	0.000	0.000	0.000
$\lambda_C(2)$	M_{AC}	-0.194	0.116	-0.426	0.027
$\lambda_{AC}(2,2)$	M_{AC}	0.000	0.000	0.000	0.000
$\lambda_{SC}(2,2)$	M_{ASC}	0.460	0.134	0.199	0.735
$\lambda_{ASC}(2,2,2)$	M_{ASC}	0.000	0.000	0.000	0.000
Model 6: $AS + SC$					
Parameter	Marginal table	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$
$\lambda_{\emptyset}$	M_{AC}	-1.418	0.025	-1.483	-1.388
$\lambda_A(2)$	M_{AC}	-0.042	0.114	-0.261	0.173
$\lambda_C(2)$	M_{AC}	-0.195	0.110	-0.414	0.020
$\lambda_{AC}(2,2)$	M_{AC}	0.000	0.000	0.000	0.000
$\lambda_S(2)$	M_{ASC}	-0.238	0.137	-0.493	0.044
$\lambda_{AS}(2,2)$	M_{ASC}	-0.291	0.137	-0.554	-0.019
$\lambda_{SC}(2,2)$	M_{ASC}	0.437	0.137	0.178	0.712
$\lambda_{ASC}(2,2,2)$	M_{ASC}	-0.086	0.143	-0.370	0.207
Model 8: ASC (Saturated)					
Parameter	Marginal table	Mean	St.dev.	$Q_{0.025}$	$Q_{0.975}$
$\lambda_{\emptyset}$	M_{ASC}	-2.325	0.079	-2.504	-2.191
$\lambda_A(2)$	M_{ASC}	-0.106	0.134	-0.379	0.152
$\lambda_S(2)$	M_{ASC}	-0.246	0.131	-0.510	0.004
$\lambda_{AS}(2,2)$	M_{ASC}	-0.292	0.139	-0.576	-0.033
$\lambda_C(2)$	M_{ASC}	-0.136	0.143	-0.402	0.151
$\lambda_{AC}(2,2)$	M_{ASC}	-0.084	0.139	-0.355	0.202
$\lambda_{SC}(2,2)$	M_{ASC}	0.450	0.135	0.207	0.705
$\lambda_{ASC}(2,2,2)$	M_{ASC}	-0.074	0.143	-0.368	0.209

Table 7: Alcohol Data

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

To save space we do not report here posterior summaries for model parameters, they can be found in a separate appendix on the web page:

`http://stat-athens.aueb.gr/~jbn/papers/paper21.htm`.

7 Discussion and Final Comments

In this paper we have dealt with the Bayesian analysis of graphical models of marginal independence for three way contingency tables. We have worked using the probability parameters of marginal tables required to fully specify each model. We have used a parameterization which directly originate from the constraints imposed by the marginal association structure of the graph. The resulting parameterization and the corresponding decomposition of the likelihood simplifies the problem and automatically imposes the marginal independences represented by the considered graph. By this way,

Table 8: Alcohol data: Posterior model probabilities and the corresponding log-marginal likelihoods; empty cell in posterior probabilities means that it is lower than 0.0001

Posterior model probabilities (%)

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

Log-marginal likelihood for each model

	Model							
	H+A+O	HA+O	HO+A	AO+H	HA+HO	HA+AO	HO+AO	HAO
Prior Distribution	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Jeffreys ($\alpha(i) = 1/2$)	-79.22	-80.11	-77.24	-87.73	-90.44	-100.93	-98.06	-98.95
UEC ($\alpha(i) = 1$)	-78.51	-78.47	-75.99	-84.70	-85.27	-93.99	-91.51	-91.46
Empirical Bayes ($\alpha(i) = p(i)$)	-86.96	-94.10	-88.94	-107.26	-124.75	-143.06	-137.91	-145.04
Perks ($a(i) = 1/\|I\|$)	-86.90	-93.19	-88.33	-107.10	-121.13	-139.89	-135.03	-141.33

the posterior model probabilities and the posterior distributions for the used parameters can be calculated analytically. Moreover, the posterior distributions of the marginal log-linear parameters $\boldsymbol{\lambda}^G$ and the probabilities $\boldsymbol{\pi}$ of the full table can be easily obtained using simple Monte Carlo schemes. This approach avoids the problem of the inverse calculation of $\boldsymbol{\pi}$ when the marginal association log-linear parameters $\boldsymbol{\lambda}$ are available which can be only achieved via an iterative procedure; see Rudas and Bergsma (2004) and Lupparelli (2006) for more details.

In the three way case all the considered models are Markov equivalent to a DAG; see Pearl and Wermuth (1994) and Drton and Richardson (2008). An immediate question which arises is whether the the graphical structures implied using the parameterization illustrated in this paper are the same with the ones that we would derive using the parameterization implied by the corresponding DAG representation. For example, in our approach the parameters for the edge model $\{AB, C\}$ with $e = C$ are given by $\boldsymbol{\pi}^G = \left(\boldsymbol{\pi}_{AB}, \boldsymbol{\pi}_C\right)$ while the parameterization implied by the corresponding DAG structure is either $\boldsymbol{\pi}^G = \left(\boldsymbol{\pi}_{A|B}, \boldsymbol{\pi}_B, \boldsymbol{\pi}_C\right)$ or $\boldsymbol{\pi}^G = \left(\boldsymbol{\pi}_{B|A}, \boldsymbol{\pi}_A, \boldsymbol{\pi}_C\right)$. The answer is given by Heckerman *et al.* (1995) where they prove that the posterior distributions and the marginal likelihoods will be the same if the priors are compatible across models since some normalizing constants cancel out. This result can be easily confirmed in the above simple example with model $\{AB, C\}$. Naturally, with our methodology we obtain also information regarding the marginal association between the variables.

An obvious extension of this work is to implement the same approach in tables of higher dimension starting from four way tables. Although most of the models in a four way contingency table can be factorized and analyzed in a similar manner, two type

of graphs (the 4-chain and the cordless four-cycle graphs) cannot be decomposed in the above way. These models are not Markov equivalent to any directed acyclic graph (DAG). In fact each bi-directed graph (which corresponds to a marginal association model) is equivalent to a DAG, i.e. a conditional association model, with the same set of variables if and only if it does not contain any 4-chain, see Pearl and Wermuth (1994). We believe that also in higher dimensional problems our approach can be applied to bi-directed graphs that admit a DAG representation. For the graph that do not factorize, more sophisticated techniques must be adopted in order to obtain the posterior distribution of interest and the corresponding marginal likelihood needed for the model comparison (work in progress by the authors).

Another interesting subject is how to obtain the posterior distributions in the case that someone prefers to work directly with marginal log-linear parameters $\boldsymbol{\lambda}^G$ defined by (4). Using our approach, we impose a prior distribution on the probability parameters $\boldsymbol{\pi}^G$. The prior of $\boldsymbol{\lambda}^G$ cannot be calculated analytically since we cannot have the inverse expression of (4) in closed form. Nevertheless, we can obtain a sample from the imposed prior on $\boldsymbol{\lambda}^G$ using a simple Monte Carlo scheme. More specifically, we can generate random values of $\boldsymbol{\pi}^G$ from the Dirichlet based prior set-ups described in this paper. We calculate the joint probability vector $\boldsymbol{\pi}$ according to the factorization of the graph under consideration and finally use (4) to obtain a sample from the imposed prior $f(\boldsymbol{\lambda}^G|G)$. This will give us an idea of the prior imposed on the log-linear parameters.

Appendix

1. **Construction of Matrix M**

   Let $\mathcal{M} = \{M_1, M_2, \ldots, M_{|\mathcal{M}|}\}$ be the set of considered marginals. Let $\mathbf{B}$ be a binary matrix of dimension $|\mathcal{M}| \times |\mathcal{V}|$ with elements B_{iv} indicating whether a variable v belongs to a specific marginal M_i. The rows of $\mathbf{B}$ correspond to the marginals in $\mathcal{M}$ whereas the columns to the variables. The variables follow a reverse ordering, that is column 1 corresponds to variable $X_{|\mathcal{V}|}$, column 2 to variable $X_{|\mathcal{V}|-1}$ and so on. Matrix $\mathbf{B}$ has elements

$$B_{iv} = \begin{cases} 1 & \text{if } v \in M_i \\ 0 & \text{otherwise.} \end{cases},$$

   for every $v \in \mathcal{V}$.

   The marginalization matrix $\mathbf{M}$ can be constructed using the following rules.

   (a) For each marginal M_i, the probability vector of the corresponding marginal table is given by $\mathbf{M}_i\pi$; where $\mathbf{M}_i$ is calculated as a Kronecker product of matrices $\mathbf{A}_{iv}$

$$\mathbf{M}_i = \bigotimes_{v \in \mathcal{V}} \mathbf{A}_{iv}$$

   with

$$\mathbf{A}_{iv} = \begin{cases} \mathbf{I}_{\ell_v} & \text{if} \quad B_{iv} = 1 \\ \mathbf{1}_{\ell_v}^T & \text{if} \quad B_{iv} = 0 \end{cases}$$

   where ℓ_v is the number of levels for v variable, $\mathbf{I}_{\ell_v}$ is the identity matrix of dimension $\ell_v \times \ell_v$ and $\mathbf{1}_{\ell_v}$ is a vector of dimension $\ell_v \times 1$ with all elements equal to one.

(b) Matrix $\mathbf{M}$ is constructed by stacking all the $\mathbf{M}_i$ matrices

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_i \\ \vdots \\ \mathbf{M}_{|\mathbb{M}|} \end{pmatrix}$$

2. **Construction of Matrix C**

Firstly we need to construct the design matrix X for the saturated model corresponding to sum to zero constraints. It has has dimension $\left(\prod_v \ell_v\right) \times \left(\prod_v \ell_v\right)$ and can be obtained as

$$\mathbf{X} = \bigotimes_{v \in \mathcal{V}^R} \mathbf{J}_{\ell_v}$$

with

$$\mathbf{J}_{\ell_v}(r,c) = \begin{cases} 1 & \text{if } c = 1 \text{ or } r = c \\ -1 & \text{if } r = 1 \text{ and } c > 1 \\ 0 & \text{otherwise.} \end{cases}$$

In matrix notation

$$\mathbf{J}_{\ell_v} = \begin{pmatrix} 1 & -\mathbf{1}^T_{(\ell_v - 1)} \\ \mathbf{1}_{(\ell_v - 1)} & \mathbf{I}_{(\ell_v - 1) \times (\ell_v - 1)} \end{pmatrix}$$

where $\mathbf{1}_{(\ell_v - 1)}$ is $(\ell_v - 1) \times 1$ vector of ones while $\mathbf{I}_{(\ell_v - 1) \times (\ell_v - 1)}$ is an identity matrix of dimension $(\ell_v - 1) \times (\ell_v - 1)$.

The contrast matrix $\mathbf{C}$ can be constructed by using the following rules.

(a) For each margin M_i construct the design matrix X_i corresponding to the saturated model (using sum to zero constraints) and invert it to get the contrast matrix for the saturated model $\mathbf{C}_i = X^{-1}$. Let $\mathbf{C}_i^*$ be a submatrix of $\mathbf{C}_i$ obtained by deleting rows not corresponding to elements of $\boldsymbol{\mathcal{E}}_{M_i}$ (the effects that we wish to estimate from margin M_i) .

(b) The contrast matrix $\mathbf{C}$ is obtained by direct sum of the C_i^* matrices as follow

$$\mathbf{C} = \bigoplus_{i:\ \mathcal{M}_i \in \mathbb{M}} \mathbf{C}_i^*$$

that is it is a block diagonal matrix with $(\mathbf{C}_1^*; M_i \in \mathcal{M})$ as the blocks. For example $C_i^* \bigotimes C_2 = \bigoplus_{i=1}^{2} C_i^*$ is the block diagonal matrix with C_1 and C_2 as blocks.

References

[1] Bartlett, M. S. (1957). A comment on Lindley's statistical paradox. *Biometrika*, **44**, 533-534.

[2] Bergsma, W. P. and Rudas, T. (2002). Marginal log-linear models for categorical data. *Annals of Statistics*, **30**, 140-159.

[3] Chen, M.H., Ibrahim, J.G. and Shao, Q. M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, **84**, 121-137.

[4] Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, **8**, 204-218, 247-277.

[5] Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980). Markov Fields and Log-Linear Interaction Models for contingency Tables. *Annals of Statistics*, **8**, 522-539.

[6] Dawid A.P. and Lauritzen S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272-1317.

[7] Dawid, A.P. and Lauritzen, S.L. (2000). Compatible prior distributions. In *Bayesian Methods with Applications to Science Policy and Official Statistics. The sixth world meeting of the International Society for Bayesian Analysis* (ed. E.I. George), 109-118. `http://www.stat.cmu.edu/ISBA/index.html`.

[8] Dellaportas, P., Forster, J.J. and Ntzoufras I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27-36.

[9] Drton, M. and Richardson, T. S. (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society B* , **70**, 287-309.

[10] Glonek, G. J. N. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society B*, **57**, 533-546.

[11] Healy, M.J.R. (1988). *Glim: An Introduction*, Claredon Press, Oxford, UK.

[12] Heckerman, D., Geiger, D. and Chickering, D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, 194- 243.

[13] Ibrahim J.G. and Chen M. H. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, **15**, 46-60.

[14] Kauermann, G. (1996). On a Dualization of Graphical Gaussian Models. *Scandinavian Journal of Statistics*, **23**, 105116.

[15] Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, **44** , 1061-1071.

[16] Lang, J. B. and Agresti, A. (1994). Simultaneously modeling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625 632.

[17] Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society B*, **54**, 340.

[18] Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187192.

[19] Lupparelli, M. (2006). *Graphical models of marginal independence for categorical variables*. Ph. D. thesis, University of Florence.

[20] Lupparelli M., Marchetti, G. M. and Bergsma, W. P. (2009). Parameterization and fitting of discrete bi-directed graph models. *Scandinavian Journal of Statistics*, **36**, 559-576

[21] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, Chapman and Hall, London.

[22] Pearl, J. and Wermuth, N. (1994). When can association graphs admit a causal interpretation? In *Models and data, artificial intelligence and statistics iv*, Cheesman P. and Oldford,W., eds.,Springer, New York, 205-214.

[23] Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the institute of actuaries*, **73**, 285-334.

[24] Richardson, T. S. (2003). Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* **30**, 145-157.

[25] Roverato A. and Consonni G. (2004) Compatible Prior Distributions for DAG models. *Journal of the Royal Statistical Society B*, **66**, 47-61.

[26] Rudas, T. and Bergsma, W. P. (2004). On applications of marginal models for categorical data. *Metron*, LXII, 1-25.

[27] Steck, H. and Jaakkola, T. (2002). On the Dirichlet Prior and Bayesian Regularization. *NIPS*, 697-704

[28] Steck, H. (2008). Learning the Bayesian Network Structure: Dirichlet Prior vs Data. *Proceedings of the conference on Uncertainty in Artificial Intelligence*, 511-518

[29] Ueno, M. (2008). Learning Likelihood-Equivalence Bayesian Networks Using an Empirical Bayesian Approach. *Behaviormetrika*, **35**, 115-135.