Quaderni di Dipartimento

Bayesian model comparison based on expected posterior priors for discrete decomposable graphical models

Guido Consonni
(Università di Pavia)

Monia Lupparelli
(Università di Bologna)

95 (05-09)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Maggio 2009

Bayesian model comparison based on expected posterior priors for discrete decomposable graphical models

Guido Consonni * and Monia Lupparelli †

Dipartimento di Economia Politica e Metodi Quantitativi
University of Pavia, Italy.

July 25, 2008

Abstract

The implementation of the Bayesian paradigm to model comparison can be problematic. In particular, prior distributions on the parameter space of each candidate model require special care. While it is well known that improper priors cannot be used routinely for Bayesian model comparison, we claim that in general the use of conventional priors (proper or improper) for model comparison should be regarded as suspicious, especially when comparing models having different dimensions. The basic idea is that priors should not be assigned separately under each model; rather they should be related across models, in order to acquire some degree of compatibility, and thus allow fairer and more robust

* *email:* guido.consonni@unipv.it

† *email:* mlupparelli@eco.unipv.it

comparisons. In this connection, the Expected Posterior Prior (EPP) methodology represents a useful tool. In this paper we develop a procedure based on EPP to perform Bayesian model comparison for discrete undirected decomposable graphical models, although our method could be adapted to deal also with Directed Acyclic Graph models. We present two possible approaches. One, based on imaginary data, requires to single-out a base-model, is conceptually appealing and is also attractive for the communication of results in terms of plausible ranges for posterior quantities of interest. The second approach makes use of training samples from the actual data for constructing the EPP. It is universally applicable, but has limited flexibility due to its inherent double-use of the data. The methodology is illustrated through the analysis of a $2 \times 3 \times 4$ contingency table.

Some key words: Bayes factor; Clique; Conjugate family; Contingency table; Decomposable model; Imaginary data; Importance sampling; Robustness; Training sample.

1 Introduction

Model comparison is an important area of Statistics. The Bayesian view is especially suited for this purpose, see for instance the review articles by George (2005) and Berger (2005). However, its implementation can be problematic, especially when comparing models having different dimensions. In particular, prior distributions on the parameter space of each model, which are required to compute Bayes factors and posterior model probabilities, need special care, because sensitivity to prior specifications in Bayesian testing and model comparison is more critical than in Bayesian inference within a single model. Specifically, conventional priors can be employed for the latter, but their use is suspicious for model comparison. The problem goes much deeper than the simple realization that improper priors cannot be naively used for computing Bayes factors, because arbitrary normalizing constants do not vanish. Indeed also proper priors are not free form difficulties when comparing hypotheses of different dimensions, as witnessed by the celebrated Jeffreys-Lindley paradox (see e.g. Robert, 2001, p. 234). The main difficulty stems from the high sensitivity of Bayes factors to the specifications of hyperparameters controlling prior-diffuseness. We claim that, when dealing simultaneously with several models, one cannot elicit priors in isolation conditionally on each single model; rather, one should take a global view and relate priors across models. This leads us straight into the area of *compatible* priors, see e.g. Dawid & Lauritzen (2001) and Consonni & Veronese (2008). In this connection, the Expected Posterior Prior (EPP) methodology of Pérez & Berger (2002) represents a useful tool. Although motivated, like the intrinsic prior approach, by the need to use objective, typically improper, priors for model choice, the EPP method has a

wider scope, and can address issues such as compatibility of priors and robustness of Bayes factors to prior elicitation. Additionally, the EPP-methodology embodies a natural tuning coefficient, the training sample size, which represents a valuable communication device to report a range of plausible values for the Bayes factor (or posterior probability) in the light of the data; see Consonni & La Rocca (2008) for an application.

This paper performs Bayesian model determination for discrete decomposable (undirected) graphical models using the EPP methodology. Specifically, Section 2 contains background material on graphical models and notation; Section 3 presents useful results originally developed by Consonni & Massam (2007): an efficient parameterization of discrete decomposable graphical models, a class of conjugate priors, as well as a reference prior. Section 4 and 5, with their specific focus on discrete graphical models, constitute the innovative part of the paper: the former develops a 'base-model', as well as an 'empirical distribution', version of Expected Posterior Prior; while the latter presents an EPP-based Bayesian model comparison methodology. Section 6 applies the methodology to a 2 x 3 x 4 contingency table representing the classification of 491 subjects according to three categorical variables, namely hypertension, obesity, and alcohol intake, with the objective of identifying the most promising models for the explanation of these data. Finally, Section 7 presents some concluding remarks.

2 Background and notation

We briefly recall some basic facts about undirected graphical models; for further details see Lauritzen (1996). Let V be a finite set of vertices; and define E to be a subset of

$V \times V$ containing unordered pairs $\{\gamma, \delta\}$, $\gamma \in V$, $\delta \in V$, $\gamma \neq \delta$. An undirected graph G is the pair (V, E). An undirected graph is *complete* if all vertices are joined by an edge. Any subset of vertices $C \subseteq V$ induces a subgraph G_C with $E_C = (C \times C) \cap E$. A subset $C \subseteq V$ is a *clique* if the induced subgraph G_C is complete. We take cliques to be maximal with respect to inclusion. In this work we focus on the class of *decomposable* undirected graphs, i.e. graphs which do not contain a chordless four cycle.

For a given ordering $C_1, \ldots, C_k$ of the cliques of a decomposable undirected graph G, we will use the following notation

$$H_l = \bigcup_{j=1}^{l} C_j,\ l = 1, \ldots, k, \quad S_l = H_{l-1} \cap C_l,\ l = 2, \ldots, k, \quad R_l = C_l \setminus S_l,\ l = 2, \ldots, k.$$

The set H_l is called the l-th *histor*y, S_l the l-th *separator* and R_l the l-th *residual*. The ordered sequence of the cliques is said to be *perfect* if for any $l > 1$ there is an $i < l$ such that $S_l \subseteq C_i$.

Given a random vector $A = (A_\gamma, \gamma \in V)$, a *graphical model*, Markov with respect to an undirected graph G, is a family of joint probability distributions on A such that $A_\delta \perp\!\!\!\perp A_\gamma \mid A_{V \setminus \{\delta, \gamma\}}$, for any pair $\{\delta, \gamma\} \notin E$. We assume A to be a discrete random vector, with each element A_γ taking values in the finite set $\mathcal{I}_\gamma$. For a given undirected decomposable graph G, we use for simplicity the same symbol G also to denote a discrete graphical model, Markov with respect to the graph G.

The Cartesian product $\mathcal{I} = \times_{\gamma \in V} \mathcal{I}_\gamma$ defines a table whose generic element

$$i = (i_\gamma, \gamma \in V)$$

is called a *cell* of the table. Consider N units, and assume that each one can be classified into one and only one of the $|\mathcal{I}|$ cells. Let $y(i)$ be the i-th cell-count; then

the collection of cell counts

$$y = (y(i), i \in \mathcal{I}), \quad \sum_{i \in \mathcal{I}} y(i) = N,$$

defines a contingency table. Conditionally on the probability $p(i)$ that a randomly chosen unit belongs to cell $i \in \mathcal{I}$, y is distributed according to a multinomial model $\mathcal{M}u(y|p, N)$, with

$$p = (p(i), i \in \mathcal{I}), \quad p(i) \geq 0, \sum_{i \in \mathcal{I}} p(i) = 1.$$

Clearly p belongs to the $|\mathcal{I}|$ dimensional simplex.

For every non-empty set $E \subseteq V$, let

$$i_E = (i_\gamma, \gamma \in E), \quad i_E \in \mathcal{I}_E = \times_{\gamma \in E} \mathcal{I}_\gamma$$

denote the cell in the E-marginal table; further denote with $p(i_E)$ and $y(i_E)$ the corresponding marginal probability and observed cell-count

$$p(i_E) = \sum_{j \in \mathcal{I} | j_E = i_E} p(j), \quad y(i_E) = \sum_{j \in \mathcal{I} | j_E = i_E} y(j).$$

For every C_l, let

$$p^{C_l} = (p(i_{C_l}), i_{C_l} \in \mathcal{I}_{C_l}), \quad y^{C_l} = (y(i_{C_l}), i_{C_l} \in \mathcal{I}_{C_l}), \quad l = 1, \ldots, k$$

be the probabilities and observed cell-counts in the C_l marginal table with $i_{C_l} = (i_{R_l}, i_{S_l})$. Let

$$p(i_{R_l}|i_{S_l}) = \frac{p(i_{C_l})}{p(i_{S_l})}, \quad l = 2, \ldots, k$$

be the probability of cell i_{R_l} conditional on cell i_{S_l}. For fixed i_{S_l} let

$$p^{R_l|i_{S_l}} = (p(i_{R_l}|i_{S_l}), i_{R_l} \in \mathcal{I}_{R_l}), \quad l = 2, \ldots, k$$

denote the vector of conditional probabilities and let

$$y^{R_l|i_{S_l}} = (y(i_{R_l}, i_{S_l}), i_{R_l} \in \mathcal{I}_{R_l}), \quad l = 2, \ldots, k$$

be the cell-counts in the C_l marginal table with a fixed configuration of i_{S_l} of S_l.

3 A parameterization and a family of prior distributions for discrete decomposable graphical models

Consonni & Massam (2007) provide several parameterizations for a decomposable undirected graphical model; additionally they derive the corresponding group-reference priors where the parameter grouping arises naturally from the graphical structure. Here we consider only one such parameterization and the allied reference prior. Whenever easily understood, probability distributions will be written without explicitly indicating their support.

3.1 Parameterization

Let G be an undirected decomposable graph, and denote with $C_1, \ldots, C_k$ the collection of its cliques arranged in a perfect ordering. Using the notation introduced in Section 2, we can write the joint multinomial distribution of counts $y = y(i)$, $i \in \mathcal{I}$, Markov with respect to G, as

$$f_G(y|p_G^{cond}, N) = \frac{N!}{\prod_{i\in\mathcal{I}} y(i)!} \prod_{i_{C_1}\in\mathcal{I}_{C_1}} \left(p(i_{C_1})\right)^{y(i_{C_1})} \prod_{l=2}^{k} \prod_{i_{S_l}\in\mathcal{I}_{S_l}} \prod_{i_{R_l}\in\mathcal{I}_{R_l}} \left(p(i_{R_l}|i_{S_l})\right)^{y(i_{c_l})}. \quad (1)$$

In the sequel, we will refer to (1) as the discrete decomposable graphical model G. The parameterization

$$p_G^{cond} = (p^{C_1}, p^{R_l|i_{S_l}}, i_{S_l} \in \mathcal{I}_{S_l}, l = 2, \ldots, k) \tag{2}$$

includes the conditional probabilities $p^{R_l|i_{S_l}}$ in the C_l marginal table, for $l = 2, \ldots, k$, as well as p^{C_1} which can also be regarded as a conditional probability upon setting $R_1 = C_1$ and $S_1 = \emptyset$. The parameter p_G^{cond} comprises $(1 + \sum_{l=2}^{k} |\mathcal{I}_{S_l}|)$ groups of parameters, each being defined on a suitable simplex. We remark that the parameterization p_G^{cond} depends on the specific perfect ordering $C_1, \ldots, C_k$.

It is expedient to rewrite the density $f_G(y|p_G^{cond}, N)$ in (1) in terms of products of multinomial densities. Let $v = (v(i),\, i \in \mathcal{I},\ \ \sum_{i\in\mathcal{I}} v(i) = L)$, and define

$$h(v|L) = \frac{L!}{\prod_{\in\mathcal{I}} v(i)!}.$$

Then

$$\begin{aligned} f_G(y|p_G^{cond}, N) &= h(y|N) \\ &\times \left(h(y^{C_1}|N)\right)^{-1} \mathcal{M}u(y^{C_1}|p^{C_1}, N) \\ &\times \prod_{l=2}^{k} \prod_{i_{S_l}\in\mathcal{I}_{S_l}} \left(h(y^{R_l|i_{S_l}}|N(i_{S_l}))\right)^{-1} \mathcal{M}u(y^{R_l|i_{S_l}}|p^{R_l|i_{S_l}}, N(i_{S_l})), \end{aligned} \tag{3}$$

with $N(i_{S_l}) = \sum_{i_{R_l}\in\mathcal{I}_{R_l}} y(i_{R_l}, i_{S_l})$. The expression $\mathcal{M}u(v|q, L)$ indicates the multinomial density with cell probabilities q and L trials evaluated in v.

We denote with G_0 the model of complete independence, under which $\perp\!\!\!\perp_{\gamma\in V} A_\gamma$. In this case: $C_\gamma = \gamma$ (for all $\gamma \in V$), $S_\gamma = \emptyset$ and $R_\gamma = C_\gamma$ $(\gamma \in V)$. The corresponding parameterization is based on marginal probabilities, but for notational coherence we still write $p_{G_0}^{cond} = (p^\gamma, \gamma \in V)$.

When the sampling distribution of $v = (v(i),\, i \in \mathcal{I},\ \ \sum_{i\in\mathcal{I}} v(i) = L)$ is standard multinomial, and the prior on the cell-probabilities p is a conjugate Dirichlet distribution, $\mathcal{D}i(p|\alpha)$, with hyperparameter α, it is well known that the marginal distribution of v is multinomial-Dirichlet, written $\mathcal{M}u\mathcal{D}i(v|\alpha, L)$, (see e.g. Bernardo & Smith, 1994, p.135). For later use we report its expression in the Appendix. We now extend this classic result to a discrete decomposable graphical model G.

Lemma 3.1. *Let the sampling distribution of y be a discrete decomposable graphical model G, so that the joint density of y, conditionally on p_G^{cond}, is given by (3). Let the prior distribution on p_G^{cond} be conjugate, namely*

$$\pi_G^C(p_G^{cond}|\alpha^G) = \mathcal{D}i(p^{C_1}|\alpha^{C_1}) \times \prod_{l=2}^{k} \prod_{i_{S_l}\in\mathcal{I}_{S_l}} \mathcal{D}i(p^{R_l|i_{S_l}}|\alpha^{R_l|i_{S_l}}), \tag{4}$$

where $\alpha^G = (\alpha^{C_1}, \alpha^{R_l|i_{S_l}}, i_{S_l} \in \mathcal{I}_{S_l}, l = 2, \ldots, k)$. Then, the marginal distribution of y is

$$\begin{aligned} m_G^C(y|\alpha^G) &= h(y|N) \\ &\times \big(h(y^{C_1}|N)\big)^{-1} \mathcal{M}u\mathcal{D}i(y^{C_1}|\alpha^{C_1}, N) \\ &\times \prod_{l=2}^{k} \prod_{i_{S_l}\in\mathcal{I}_{S_l}} \big(h(y^{R_l|i_{S_l}}|N(i_{S_l})\big)^{-1} \mathcal{M}u\mathcal{D}i(y^{R_l|i_{S_l}}|\alpha^{R_l|i_{S_l}}, N(i_{S_l})). \end{aligned} \tag{5}$$

The proof is trivial, since the computation reduces to a collection of independent standard multinomial-Dirichlet problems. Because of conjugacy, the posterior distribution of p_G^{cond} also belongs to the family (4) with updated parameters

$$\alpha^{C_1} \to \alpha^{C_1} + y^{C_1}, \quad \alpha^{R_l|i_{S_l}} \to \alpha^{R_l|i_{S_l}} + y^{R_l|i_{S_l}}.$$

As a consequence, result (5) immediately provides also the expression of the predictive distribution.

3.2 Reference prior

Reference analysis provides one of the most successful general methods to derive default prior distributions on multidimensional parameters. For a recent and informative review, see Bernardo (2005). For an application to a multinomial setting, see Berger & Bernardo (1992). The definition of a reference prior applies to a specific parameterization; moreover it requires the user to specify groups of parameters and an ordering of inferential importance of the groups.

The $(1+\sum_{l=2}^{k}|\mathcal{I}_{S_l}|)$ group-reference prior for p_G^{cond}, with parameter groupings identified in (2), is

$$\pi_G^R(p_G^{cond}) \propto \Big(\prod_{i_{c_1}\in\mathcal{I}_{C_1}} p(i_{C_1})\Big)^{-\frac{1}{2}} \prod_{l=2}^{k}\prod_{i_{S_l}\in\mathcal{I}_{S_l}} \Big(\prod_{i_{R_l}\in\mathcal{I}_{R_l}} p(i_{R_l}|i_{S_l})\Big)^{-\frac{1}{2}}. \tag{6}$$

For the derivation of (6) and further properties, see Consonni & Massam (2007). In particular they show that the order of the groupings is irrelevant. Notice that the reference prior is proper, being a product of Jeffreys' priors, one for each of the groups of p_G^{cond}, which are thus *a priori* independent. This structural property corresponds to the notion of *global* and *local* independence, introduced by Geiger & Heckerman (1997) for the analysis of Directed Acyclic Graphs. It is reassuring that in the p_G^{cond} parameterization this useful property does not arise out of convenience but actually stems from applying the reference prior algorithm. The reference prior (6) is clearly conjugate since it belongs to the family (4) with α^{C_1} and $\alpha^{R_l|i_{S_l}}$ each being $\frac{1}{2}\underline{1}$, where $\underline{1}$ denotes a vector of 1's of suitable dimension. Accordingly, the posterior also belongs to (4), and the marginal, and predictive data distribution can be obtained as a special case of (5). For clarity we will later use the superscript 'R' to remind the reader that we are using the reference prior (6) instead of a subjectively specified conjugate prior.

4 Expected posterior priors for decomposable discrete graphical models

4.1 General

Bayesian model comparison typically requires the computation of marginal data distribution in order to derive Bayes Factors (BF's) and ultimately posterior model probabilities. As recalled in the Introduction, improper priors cannot be routinely used, and this has led to an active research in the area of objective model comparison, see Berger & Pericchi (2001), and Pericchi (2005). We have also stressed that in general conventional priors on the parameter space of each model are problematic, and that sensitivity is a pervasive issue.

The *Expected Posterior Prior* (EPP) approach developed by Pérez & Berger (2002), represents a useful tool to address the above difficulties. This method is similar to that of 'information transfer' between models, originally proposed by Neal (2001). It also bears strong connection to the intrinsic prior methodology, see e.g. Berger & Pericchi (1996), and indeed the EPP for the pairwise comparison of two nested models actually coincides with the intrinsic prior for that problem.

A basic idea in the EPP approach is to make use of *imaginary* data, which has been for a long time a useful ingredient in Bayesian thinking. Consider model $\mathcal{M}_k$, with parameter θ_k, and let $\pi_k^N(\theta_k)$ be a default, noninformative, possibly improper prior for θ_k. Suppose that x represent imaginary observations, and let $m^*(x)$ be a suitable *marginal*, distribution for x. The smallest x inducing a proper 'posterior' $\pi_k^N(\theta_k|x)$ constitutes a *minimal training sample*. (Notice that we use the term 'training' also

when referring to imaginary data). We are now ready to define the EPP for θ_k, relative to m^*, as

$$\pi_k^*(\theta_k) = \int \pi_k^N(\theta_k|x) m^*(x) dx.$$

Notice that the EPP is an average of fictitious posteriors with respect to the chosen marginal m^*. Provided m^* satisfies some weak requirements, the EPP enjoys several advantages. We mention here four of them. Firstly indeterminate constants possibly present in the π_k^N's disappear when computing the BF's (this is trivial because $\pi_k^N(\theta_k|x)$ is assumed to be proper). Secondly, if m^* is proper, then π_k^* is also proper; conversely, if m^* is not proper, indeterminacy of the resulting BF will *not* arise, because the same m^* is used for all models, and thus again arbitrary normalizing constants cancel out. Thirdly, notice that this method only requires one distribution to be elicited, namely m^*, since all the remaining priors π_k^N are, by assumption, automatically assigned according to some default technique. Finally, all priors $\pi_k^*(\theta_k)$, $\theta_k \in \Theta_k$, achieve some sort of compatibility among themselves, since each is 'shrunk' on a subregion of Θ_k which is consistent with the mixing distribution $m^*(x)$.

As for the choice of m^*, a few proposals have been put forward, which we briefly recollect here. Suppose one can identify a base-model $\mathcal{M}_0$ for a given problem; this is usually the simplest possible model (e.g. in variable selection that having only the intercept). Then the marginal data distribution under this model is a natural candidate, i.e. $m^*(x) = m_0(x)$, where $m_0(x) = \int_{\Theta_0} f_0(x|\theta_0)\pi_0^N(\theta_0)$. We call this method the base-model EPP.

When the above strategy is not feasible, a natural alternative is to set $m^*(x)$ equal to the *empirical* distribution which, for given data $y_1, \ldots, y_N$, is defined by

$m^{emp}(x) = \frac{1}{L}\sum_l I_{\{y(l)\}}(x)$, where I_S denotes the indicator function of the set S, and $y(l) = (y_{l_1}, \ldots, y_{l_M})$ is a subsample of given size M, $0 \leq M \leq N$, such that $\pi_k^N(\theta_k|y(l))$ exists for all models, and L is the number of all such samples of size M. The corresponding method is called the empirical EPP. Notice that this approach implies a double use of the data: to construct priors and to derive BF's. As a consequence, $y(l)$ is required to be a minimal training sample; for further details see Berger & Pericchi (2004).

We now detail the EPP methodology in the context of discrete graphical models.

Consider a given set of discrete decomposable graphical models $G_0, \ldots, G_U$ with G_u parameterized according to $p_{G_u}^{cond}$, as defined in (2). We let x denote an imaginary contingency table of size M having the same structure as the actual data y, so that $x = x(i)$, with $i \in \mathcal{I}$. Let $\sum_{i\in\mathcal{I}} x(i) = M$ be the training sample size. We assume that the default prior on $p_{G_u}^{cond}$ is given by the reference prior $\pi_{G_u}^R(p_{G_u}^{cond})$, see (6). We are now ready to define the EPP for $p_{G_u}^{cond}$ with respect to the marginal data distribution $m^*(x)$.

Proposition 4.1. *Given a discrete decomposable graphical model G_u, with prior $\pi_{G_u}^R(p_{G_u}^{cond})$, the corresponding EPP for $p_{G_u}^{cond}$ is*

$$\pi_{G_u}^*(p_{G_u}^{cond}) = \sum_{x:\sum_i x(i)=M} \pi_{G_u}^R(p_{G_u}^{cond}|x) m^*(x).$$

Notice that, while the groups of the $p_{G_u}^{cond}$ parameterization are independent under the reference prior $\pi_{G_u}^R(p_{G_u}^{cond})$, this is no longer so under the EPP $\pi_{G_u}^*(p_{G_u}^{cond})$. The marginal data distribution under G_u induced by the EPP can be shown to be

$$m_{G_u}^*(y) = \sum_{x:\sum_i x(i)=M} m_{G_u}^R(y|x) m^*(x). \tag{7}$$

4.2 Base-model EPP

In our context, a natural choice for the base-model is represented by the complete independence model G_0. We therefore set

$$m^*(x) = m^R_{G_0}(x) = \int f_{G_0}(x|p^{cond}_{G_0}, M)\pi^R_{G_0}(p^{cond}_{G_0})dp^{cond}_{G_0}.$$

Accordingly, the base-model EPP for model G_u is

$$\pi^{*0}_{G_u}(p^{cond}_{G_u}) = \sum_{x:\sum_i x(i)=M} \pi^R_{G_u}(p^{cond}_{G_u}|x)m^R_{G_0}(x), \tag{8}$$

with marginal data distribution

$$m^{*0}_{G_u}(y) = \sum_{x:\sum_i x(i)=M} m^R_{G_u}(y|x)m^R_{G_0}(x). \tag{9}$$

For later use, we report the analytic expression of $m^R_{G_0}(x)$ in the Appendix.

4.3 Empirical EPP

This strategy requires some careful thinking in our case because the original definition of empirical EPP is based on subsamples of *individual* observations, while our problem is more naturally cast in terms of contingency tables.

We start by considering a realized sample of N *individuals* $z = (z_1, \ldots, z_N)$, with $z \in \mathcal{Z}$. Each z_j is a $|V|$-dimensional vector whose γ-component takes values in the set of configurations of the discrete random variable A_γ, $\gamma \in V$. The N individuals can be subsequently classified in a $|V|$-dimensional contingency table $y = y(i)$, $i \in \mathcal{I}$ of size N where

$$y(i) = \sum_{j=1}^{N} I_{z_j}(i), \quad i \in \mathcal{I}.$$

Consider now the subspace $\tilde{Z}_M = \{\tilde{z}_1, \ldots, \tilde{z}_B\}$ of all subsamples of size $M \leq N$ from $z = (z_1, \ldots, z_N)$. The generic element of $\tilde{Z}_M$ is $\tilde{z}_b = (z_{b_1}, \ldots, z_{b_M})$, with $b_1, \ldots, b_M$

denoting distinct indices in $\{1, \ldots, N\}$. Let $B = \binom{N}{M}$ be the cardinality of $\tilde{\mathcal{Z}}_M$, i.e. the number of all subsamples of size M. Then, on the basis of the empirical distribution, the probability of each $\tilde{z}_b$ is $k(\tilde{z}_b) = 1/B$. Each subsample $\tilde{z}_b$ can be classified in a $|V|$-dimensional contingency table of size M where

$$\tilde{y}_{\delta(b)}(i) = \sum_{j=1}^{M} I_{\tilde{z}_{b_j}}(i), \quad i \in \mathcal{I}.$$

Notice that distinct subsamples $\tilde{z}_b$ may give rise to the same contingency table. Let $\Delta = \{1, \ldots, D\}$ represent the set of all distinct contingency tables. The contingency table generated by a subsample $\tilde{z}_b \in \tilde{Z}_M$ will be denoted by $\tilde{y}_d = (\tilde{y}_d(i), i \in \mathcal{I})$, with $d = \delta(b) \in \Delta$.

Lemma 4.1. *Let y be a contingency table of size N. The space of contingency tables generated by all subsamples of size $M \leq N$ is defined by $\tilde{y}_d = (\tilde{y}_d(i), i \in \mathcal{I})$, $d = 1, \ldots, D$, such that*

(i) $\sum_{i \in \mathcal{I}} \tilde{y}_d(i) = M$

(ii) $\tilde{y}_d(i) \leq y(i)$, *for every* $i \in \mathcal{I}$.

Under the empirical distribution, the probability of each $\tilde{y}_d$ is

$$q(\tilde{y}_d) = \frac{1}{B} \prod_{i \in \mathcal{I}} \binom{y(i)}{\tilde{y}_d(i)}, \tag{10}$$

with $\prod_{i \in \mathcal{I}} \binom{y(i)}{\tilde{y}_d(i)}$ denoting the number of subsamples $\tilde{z}_b$ having as image the same contingency table $\tilde{y}_d$ of size M. Clearly, if $M = N$, then $q(y) = 1$, where y is the observed contingency table. Therefore the empirical marginal distribution on the space of all contingency tables x of size M is given by

$$m^{emp}(x) = \sum_{d=1}^{D} I_{\tilde{y}_d}(x) q(\tilde{y}_d).$$

Setting $m^*(x) = m^{emp}(x)$, we obtain the empirical EPP for model G_u

$$\pi^{*emp}_{G_u}(p^{cond}_{G_u}) = \sum_{d=1}^{D} \pi^{R}_{G_u}(p^{cond}_{G_u}|\tilde{y}_d)q(\tilde{y}_d).$$

The marginal data distribution for any model G_u under the empirical EPP is therefore

$$m^{*emp}_{G_u}(y) = \sum_{d=1}^{D} m^{R}_{G_u}(y|\tilde{y}_d)q(\tilde{y}_d).$$

Although seemingly different form their base-model counterpart, both $\pi^{*emp}_{G_u}(p^{cond}_{G_u})$ and $m^{*emp}_{G_u}(y)$ can be written in the same format as (8), respectively (9). For example we can write $m^{*emp}_{G_u}(y)$ as

$$m^{*emp}_{G_u}(y) = \sum_{x:\sum_i x(i)=M} m^{R}_{G_u}(y|x)m^{emp}(x).$$

5 Bayesian model comparison based on EPP

Consider a collection of $G_0, \ldots, G_U$ of discrete decomposable graphical models. For every pair G_u, G_v, with $u, v = 0, \ldots, U$ and $u \neq v$, the Bayes factor based on the EPP is

$$BF^{*}_{G_uG_v}(y) = \frac{m^{*}_{G_u}(y)}{m^{*}_{G_v}(y)},$$

where $m^{*}_{G_u}(y)$ is defined in (7). The posterior probability of model G_u is then given by

$$Pr^*(G_u|y) = \Big(1 + \sum_{v\neq u} \frac{w_v}{w_u} BF^{*}_{G_vG_u}(y)\Big)^{-1}, \quad u = 0, \ldots, U, \tag{11}$$

where $w_u = Pr(G_u)$ is the prior probability of model G_u. If prior odds on model space are all equal to 1, so that $w_u/w_v = 1$ for all $u \neq v$, then formula (11) is simply a function of the Bayes factors $BF^{*}_{G_vG_u}(y)$.

5.1 Model comparison under the base-model expected posterior prior

Consider two decomposable undirected graphical models G_u and G_v. The Bayes factor calculated with respect to the base-model EPP is

$$BF^{*0}_{G_uG_v}(y) = \frac{m^{*0}_{G_u}(y)}{m^{*0}_{G_v}(y)}, \tag{12}$$

where $m^{*0}_{G_u}(y)$ is defined in (9). Pérez & Berger (2002) show that the Bayes factor in (12) satisfies the coherence condition

$$BF^{*0}_{G_vG_u}(y) = BF^{*}_{G_vG_0}(y)BF^{*}_{G_0G_u}(y),$$

where

$$BF^{*0}_{G_uG_0}(y) = \frac{m^{*0}_{G_u}(y)}{m^{R}_{G_0}(y)}. \tag{13}$$

The denominator in (13) is the marginal data distribution under model G_0 and prior $\pi^R_{G_0}(p^{cond}_{G_0})$ reproduced in the Appendix. Recall that

$$m^{*0}_{G_u}(y) = \sum_{x:\sum_i x(i)=M} m^R_{G_u}(y|x)m^R_{G_0}(x), \tag{14}$$

with $m^R_{G_u}(y|x)$ denoting the 'predictive' distribution under G_u (notice that we actually predict real data y conditionally on imaginary data x), when the prior is $\pi^R_{G_u}(p^{cond}_{G_u})$.

The marginal distribution $m^{*0}_{G_u}(y)$ requires summing over all possible contingency tables such that $\sum_i x(i) = M$. This computation can be very demanding, and virtually unfeasible, even when M and/or the dimension of the table $|\mathcal{I}|$ are only moderately large. We therefore approximate expression (14) using a Monte Carlo sum; for a similar strategy in a related context, see Casella & Moreno (2007). In particular we use an

importance sampling algorithm with the following importance function

$$g_{G_u}(x) = \mathcal{M}u(x|\hat{p}_{G_u}, M),$$

i.e. a multinomial distribution with M trials and cell-probabilities $\hat{p}_{G_u}$, the maximum likelihood estimate of p under model G_u. If we draw T independent samples $x^{(t)}$ from $g_{G_u}(x)$, the Bayes factor (13) can be approximated as

$$\widehat{BF}^{*0}_{G_u G_0}(y) = \frac{1}{m^R_{G_0}(y)} \frac{1}{T} \sum_{t=1}^{T} \frac{m^R_{G_u}(y|x^{(t)}) m^R_{G_0}(x^{(t)})}{g_{G_u}(x^{(t)})}.$$

The estimated posterior probability of model G_u is then

$$\widehat{Pr}^{*0}(G_u|y) = \left(1 + \sum_{v \neq u} \frac{w_v}{w_u} \widehat{BF}^{*0}_{G_v G_u}(y)\right)^{-1}, \quad u = 0, \ldots, U,$$

where

$$\widehat{BF}^{*0}_{G_v G_u}(y) = \widehat{BF}^{*0}_{G_v G_0}(y) \widehat{BF}^{*0}_{G_0 G_u}(y).$$

5.2 Model comparison under the empirical expected posterior prior

Given two decomposable graphical models G_u and G_v, the Bayes factor based on the empirical EPP is

$$\begin{aligned} BF^{*emp}_{G_u G_v}(y) &= \frac{m^{*emp}_{G_u}(y)}{m^{*emp}_{G_v}(y)} \\ &= \frac{\sum_{d=1}^{D} m^R_{G_u}(y|\tilde{y}_d) q(\tilde{y}_d)}{\sum_{d=1}^{D} m^R_{G_v}(y|\tilde{y}_d) q(\tilde{y}_d)}, \end{aligned} \tag{15}$$

where $q(\tilde{y}_d)$ is defined in (10). As in the previous subsection, the number of terms in both the sums of (15) can be prohibitively large. Accordingly, we propose to approximate each of the two marginal distributions appearing in (15) using an importance

sampling strategy. Then,

$$\widehat{m}^{*emp}_{G_u}(y) = \frac{1}{T}\sum_{t=1}^{T} \frac{m^R_{G_u}(y|x^{(t)})m^{emp}(x^{(t)})}{g_{G_u}(x^{(t)})}, \tag{16}$$

where the importance function $g_{G_u}(x)$ is again a multinomial distribution with cell-probabilities $\hat{p}_{G_u}$.

When using the empirical EPP, the training sample size M should be the smallest possible, to reduce double-counting. Notice that the support of the importance function is strictly larger than that of the empirical distribution; as a consequence, some random draws from the importance function are 'Not Subsamples' (NS) $\tilde{y}_d$ of y. In general, the percentage of NS should be small for an efficient approximation. A further reason to limit the value of M is that higher values of M will generally increase the percentage of NS. In particular, for $M = N$, the empirical distribution degenerates on the observed vector y, and thus the percentage of NS is likely to be close to 100. On the other hand, when M is too small, cells with low probabilities under $\hat{p}_{G_u}$ are likely to receive zero-counts in many draws, so that the importance function will result in a poor approximation to the empirical distribution. Since $\hat{p}_{G_u}$ varies across models, it would seem reasonable to adapt the minimal training sample to each specific model, setting for instance $M_{G_u} = 1/min(\hat{p}_{G_u})$. In this way, the expected count of each cell, under the importance function, is at least one, thus providing a better approximation to the empirical distribution, since the number of cell having zero-count draws are likely to be very limited. A drawback of this procedure is that it may produce highly variable values for M_{G_u} across models. In this way, model choice could be unduly driven by a different use of the data set information. To overcome this difficulty, we recommend using the same value of M for each model, and to perform some sensitivity

		Alcohol intake			
Obesity	Hypertension	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

Table 1: Alcohol, hypertension and obesity data. Alcohol intake is measured by number of drinks/day.

analysis over the range $\underline{M} \leq M \leq \overline{M}$, where $\underline{M} = \min_u\{M_{G_u}\}$ and $\overline{M} = \max_u\{M_{G_u}\}$.

The approximate Bayes factor based on the empirical EPP is

$$\widehat{BF}^{*emp}_{G_uG_v}(y) = \frac{\widehat{m}^{*emp}_{G_u}(y)}{\widehat{m}^{*emp}_{G_v}(y)},$$

where both numerator and denominator are defined in (16), leading to the approximate posterior probability of model G_u

$$\widehat{Pr}^{*emp}(G_u|y) = \Big(1 + \sum_{v\neq u} \frac{w_v}{w_u}\widehat{BF}^{*emp}_{G_vG_u}(y)\Big)^{-1}, \quad u, v = 0, \ldots, U,\ u \neq v.$$

6 Example: hypertension, obesity, and alcohol intake data

We consider the data set in Table 1 representing the classification of 491 subjects according to three categorical variables, namely hypertension (H: yes, no), obesity (O: low, average, high) and alcohol intake (A: 0, 1-2, 3-5, 6+ drinks per day). This $2 \times 3 \times 4$ table was analyzed by Knuiman & Speed (1988) and from a Bayesian model determination perspective by Dellaportas & Forster (1999).

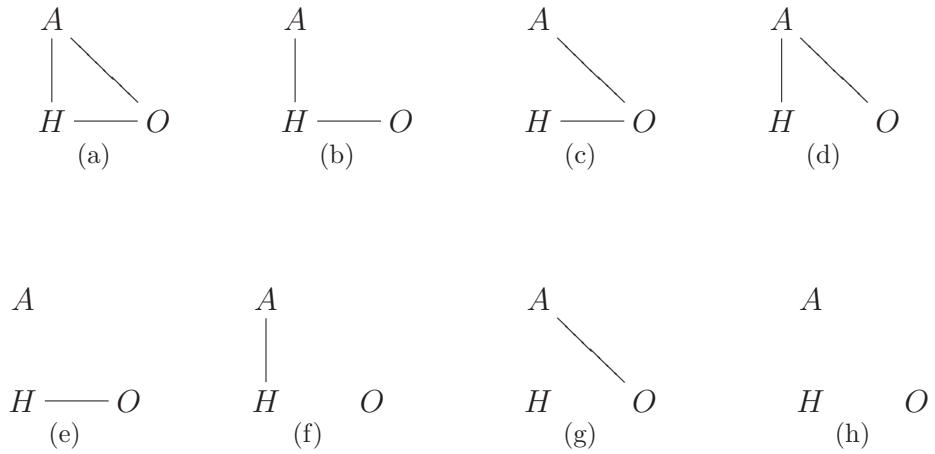

Figure 1: Top panel: undirected decomposable graphical models of conditional independence. (a) AHO: the unrestricted model. (b) $AH + HO$: $A \perp\!\!\!\perp O \mid H$. (c) $AO + HO$: $A \perp\!\!\!\perp H \mid O$. (d) $AH + AO$: $H \perp\!\!\!\perp O \mid A$. Bottom panel: undirected decomposable graphical models of marginal independence. (e) $A + HO$: $A \perp\!\!\!\perp HO$. (f) $O + AH$: $O \perp\!\!\!\perp AH$. (g) $H + AO$: $H \perp\!\!\!\perp AO$. (h) $A + H + O$: $A \perp\!\!\!\perp H \perp\!\!\!\perp O$; the complete independence model.

Altogether there exist eight possible decomposable undirected graphical models for this problem illustrated in Figure 1. We first consider a conventional approach, i.e. assigning the conjugate family of priors (4) under each model, and computing the corresponding Bayes factor

$$BF_{G_u G_v}(y|\alpha^{G_u}, \alpha^{G_v}) = \frac{m^C_{G_u}(y|\alpha^{G_u})}{m^C_{G_v}(y|\alpha^{G_v})},$$

where $m^C_{G_u}(y|\alpha^{G_u})$ is reported in (5). In particular we choose three distinct sets of values for $\alpha^{G_u} = \alpha$, namely: $\alpha_U = \underline{1}$, corresponding to a product of uniform priors; $\alpha_J = \frac{1}{2}\underline{1}$, a product of Jeffreys priors; $\alpha_P = (\frac{1}{|\mathcal{I}_{R_l}|}, l = 1, \ldots, k)$, a product of priors each corresponding to a Dirichlet distribution originally proposed by Perks, and discussed also by Dellaportas & Forster (1999). (Recall that $R_1 = C_1$; furthermore the unit vector $\underline{1}$ has variable dimension across models, but for simplicity we omit such

Model	α_U	α_J	α_P
AHO	0.000	0.000	0.000
$AH + HO$	0.344	0.146	0.004
$AO + HO$	0.001	0.000	0.000
$AH + AO$	0.000	0.000	0.000
$A + HO$	0.549	0.683	0.191
$O + AH$	0.044	0.043	0.001
$H + AO$	0.000	0.000	0.000
$A + H + O$	0.062	0.128	0.804

Table 2: Posterior model probabilities under conjugate priors, for distinct choices of α.

dependence in the notation). These priors can be thought of as being progressively more diffuse. Indeed their 'overall precision', as measured by the sum of the elements of α, is for each given model greatest under α_U, intermediate under α_J and least under α_P. For instance, each of the uniform priors on a specific residual-table has an overall precision equal to the number of cells in that subtable, while the overall precision under each of the Percks prior is equal to one.

Table 2 contains the posterior probabilities for each decomposable graphical model using conjugate priors (we assume that all prior odds are equal to one). We notice that the posterior probability is essentially concentrated on three models, namely $(AH + HO)$, $(A + HO)$ and $(A + H + O)$. However the distribution of these probabilities is highly sensitive to the value of α. Specifically, model $(A + HO)$ receives the highest posterior probability under α_J, and α_U. On the other hand, the top model under α_P is by far the independence model $(A + H + O)$. The results in columns α_J and α_p of Table 2 are very close to those obtained by Dellaportas & Forster (1999), using a hyper-Dirichlet prior, (see their Table 1). Notice however that they considered, in addition to the eight decomposable graphical model listed above, also the hierarchical non-graphical model $(HO + AH + AO)$ which however received negligible

posterior probability in all their experiments. Furthermore, their analysis with the hyper-Dirichlet involves only two priors under the unrestricted model AHO, namely a Jeffreys and a Percks Dirichlet. The results in Table 2 are also comparable to those obtained using the method suggested by Raftery (1996) and implemented in the S-plus function 'glib', again reported in Table 1 of Dellaportas & Forster (1999). These Authors also report the results, based on a particular normal prior on the log-linear parameters (using three distinct sets of variances to gauge sensitivity), obtained using a reversible-jump MCMC procedure. The latter method is shown to be less sensitive to the choice of the prior hyperparameters in the sense that the best model is that of mutual independence $(A + H + O)$, regardless of the prior variances (posterior probabilities vary in the range $51\% - 81\%$), followed by the model of independence of alcohol from the pair (hypertension-obesity), $(A + HO)$, whose posterior probabilities vary between 47% and 19%. For the choice α_U (uniform prior on the simplex under each residual-table), Table 2 reveals that the model of mutual independence is not the most likely one, receiving a bare 6% probability; much stronger evidence in given to models $(A + HO)$ and $(AH + HO)$, the latter being a model of conditional independence (between alcohol intake and obesity given hypertension). Although quite simple, this example brings home the message that Bayesian model determination is particularly sensitive to specifications regulating the degree of diffuseness of the prior, whose impact would typically be modest in conventional prior-to-posterior analysis within a single model.

We now consider model determination, for the same data set, using the base-model EPP. For each model the starting distribution was the same as the one considered in

	α_U				α_J				α_P			
Model	M_{25}	M_{50}	M_{75}	M_{100}	M_{25}	M_{50}	M_{75}	M_{100}	M_{25}	M_{50}	M_{75}	M_{100}
AHO	0.000	0.002	0.005	0.012	0.000	0.001	0.004	0.065	0.000	0.001	0.003	0.007
$AH + HO$	0.707	0.747	0.697	0.685	0.679	0.748	0.730	0.581	0.674	0.754	0.738	0.658
$AO + HO$	0.011	0.027	0.071	0.063	0.009	0.024	0.042	0.070	0.007	0.023	0.031	0.046
$AH + AO$	0.001	0.005	0.011	0.018	0.001	0.005	0.011	0.018	0.001	0.005	0.009	0.017
$A + HO$	0.234	0.174	0.162	0.158	0.262	0.178	0.159	0.196	0.270	0.175	0.168	0.174
$O + AH$	0.034	0.035	0.043	0.050	0.036	0.035	0.042	0.054	0.034	0.034	0.040	0.081
$H + AO$	0.001	0.001	0.003	0.005	0.000	0.001	0.003	0.005	0.000	0.001	0.002	0.004
$A + H + O$	0.011	0.008	0.009	0.010	0.013	0.008	0.009	0.011	0.014	0.007	0.009	0.013

Table 3: Posterior model probabilities under the base-model EPP, for different training sample sizes M, and distinct choices of α.

the conventional analysis, namely a conjugate prior (4) with three choices for α, namely α_U, α_J and α_P. Table 3 collects the results obtained according to the three values of α and the training sample size M. For the latter, four values were chosen corresponding to 25%, 50%, 75% and 100% of the actual sample size N, in order to evaluate the sensitivity of the analysis. Relative to the conventional analysis described earlier, two features emerge clearly. For each fixed M there is now a broad agreement between the results obtained under the three distinct priors; in particular the behaviour under α_P is now comparable to the other choices of α: in other words robustness with respect to the starting prior has been achieved. Secondly, the highest posterior probability model is now $(AH + HO)$ followed by $(A + HO)$ (notice the interchange of ranking relative to the conventional approach under α_U and α_J). Finally variation of the results with respect to M is limited, especially for the top model and if one removes the rather unrealistic case M_{100}, corresponding to a training sample size equal to the actual sample size $(M = N)$. As a further check, we have also run the analysis modifying the (perfect) ordering of the cliques, for those models which would allow alternative

	α_U		α_J		α_P^1		α_P^2	
Model	$M = 49$	$M = 101$	$M = 49$	$M = 101$	$M = 49$	$M = 101$	$M = 49$	$M = 101$
AHO	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$AH + HO$	0.625	0.806	0.561	0.774	0.000	0.000	0.484	0.585
$AO + HO$	0.003	0.009	0.002	0.007	0.002	0.023	0.002	0.011
$AH + AO$	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000
$A + HO$	0.324	0.155	0.385	0.190	0.866	0.846	0.447	0.350
$O + AH$	0.032	0.024	0.031	0.022	0.067	0.103	0.034	0.042
$H + AO$	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
$A + H + O$	0.016	0.005	0.021	0.005	0.065	0.027	0.033	0.011

Table 4: Posterior model probabilities under the empirical EPP, for different training sample sizes M, and distinct choices of α.

orderings, e.g. $(AH + HO)$. Notice that this induces a distinct parameterization p_G^{cond}, and distinct priors. However, the results were quite comparable to the ones reported in Table 3, and accordingly we omit details. Relative to the base-model EPP analysis, we can therefore conclude that the top model is $(AH + HO)$ with a posterior probability of the order of 70%, followed by model $(A + HO)$, and that these results are robust with respect to the starting model priors, as well as to the training sample size.

We finally turn to the empirical EPP analysis whose results are presented in Table 4. Again, we report posterior model probabilities under the three choices α_U, α_J and α_P and based on two different training sample sizes, namely $\underline{M} = 49$ and $\overline{M} = 101$ corresponding respectively to the model $(A + H + O)$ and $(AH + HO)$. Although not shown, we also computed the percentage of 'Not Subsamples' (NS): this is close to zero under $\underline{M}$, while it varies from almost zero to 4% under $\overline{M}$. Looking at columns α_U and α_J, we notice that the highest probability model is still $(AH + HO)$ followed by $(A + HO)$; this result is consistent with that obtained under the base-model EPP, also in terms of actual probability-values. We also verified that changing the clique ordering

(when this is applicable) does not modify the results under α_U and α_J. This conclusion does not hold however under the α_P choice. Specifically, for one ordering, which corresponds to α_P^2 in the Table 4, results are broadly similar to those just described for α_U and α_J; on the other hand, for an alternative clique-ordering, corresponding to α_P^1 in the Table, results differ. Essentially, the probability assigned to the union of the two highest posterior probability models $(AH+HO)\cup(A+HO)$ is now concentrated onto the simpler model $A+HO$. In conclusion, the EPP based on the empirical distribution confirms the finding that the two best models are $(AH + HO)$ and $(A + HO)$, with $(AH + HO)$ receiving higher probability, except for the choice of α_P^1.

7 Concluding remarks

In this paper we have developed a methodology based on Expected Posterior Priors (EPP) to perform Bayesian model comparison for discrete decomposable graphical models. In this connection, the parameterization and priors presented in Consonni & Massam (2007) proved to be particularly useful. Our method could be adapted to Directed Acyclic Graph (DAG) models, see e.g. Cowell *et al.* (1996), which however require an ordering of the variables involved. The basic idea is to replace cliques with individual nodes, and use the collection of parents instead of the separators. This would result in a new p_G^{cond} parameterization, and corresponding conjugate family of priors, which would retain the basic feature of local and global independence as described in this paper.

We have illustrated our methodology analysing a $2\times3\times4$ contingency table. Since the number of variables involved was very limited, we were able to individually con-

sider all the eight possible decomposable models, thus illustrating fully the sensitivity of a conventional analysis, as well as highlighting the main features of our method. Despite being a small-scale problem, computation of relevant quantities, such as the marginal data distribution under the EPP, required an importance sampling strategy to evaluate a sum of terms over the space of all $2 \times 3 \times 4$ contingency tables. Clearly, for problems involving a high number of variables, exhaustive consideration of each single decomposable model would be unfeasible. In this case our method could still be useful, but should be coupled with MCMC techniques to search over model space. We have discussed a base-model, as well as an empirical distribution, approach to EPP. The former presents several comparative advantages: it uses only imaginary data, thus making no double use of the actual data; as a consequence the full range of training sample sizes can be used ($0 \leq M \leq N$), so that robustness issues can be more adequately evaluated. Furthermore, as revealed by the analysis of our data set, the base-model EPP method showed no particular bias in favour of the simpler models, while exhibiting greater stability to prior specifications than the empirical distribution EPP. However, when a base-model cannot be identified for the problem at hand, the empirical EPP approach may represent a viable alternative.

Acknowledgments

Work partially supported by MIUR (PRIN 2005132307) and the University of Pavia. We thank Luca La Rocca (University of Modena and Reggio Emilia, Italy) for useful discussions. We are also grateful to Giovanni M. Marchetti (University of Florence), who supplied some R functions used in our computational analysis.

A Appendix

1. The vector $v = (v(i), i \in \mathcal{I}, \sum_{i\in\mathcal{I}} v(i) = L)$ is distributed according to the multinomial-Dirichlet family, $\mathcal{M}u\mathcal{D}i(v|\alpha, L)$, if its density is

$$m(v|\alpha) = \frac{L!}{\prod_{i\in\mathcal{I}} v(i)!} \frac{\Gamma(\alpha_+)}{\prod_{i\in\mathcal{I}} \Gamma(\alpha(i))} \frac{\prod_{i\in\mathcal{I}} \Gamma(\alpha(i) + v(i))}{\Gamma(L + \alpha_+)},$$

   where $\alpha_+ = \sum_{i\in\mathcal{I}} \alpha(i)$.

2. The marginal distribution of a set x of imaginary data of size M under model G_0 is

$$m^R_{G_0}(x|\alpha^{G_0}) = h(x|M) \prod_{\gamma\in V} \big(h(x^\gamma|M)\big)^{-1} \mathcal{M}u\mathcal{D}i(x^\gamma|\alpha^\gamma, M).$$

References

Berger, J. O. (2005). Bayes factors. In C. Balakrishnan, N. Read & B. Vidakovic, eds., *Encyclopedia of statistical science*, vol. 1. pp. 378–386.

Berger, J. O. & Bernardo, J. M. (1992). Ordered group reference priors with application to a multinomial problem. *Biometrika* **79**, 25–37.

Berger, J. O. & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison (with discussion). In P. Lahiri, ed., *Model selection*. Institute of Mathematical Statistics, Beachwood, OH, pp. 135–207.

Berger, J. O. & Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *The Annals of Statistics* **32**, 841–869.

Bernardo, J. M. (2005). Reference analysis. In D. K. Dey & R. Rao, eds., *Handbook of statistics*, vol. 25. Elsevier, Amsterdam, pp. 17–90.

Bernardo, J. M. & Smith, A. E. M. (1994). *Bayesian theory*. Wiley, Chichester.

Casella, G. & Moreno, E. (2007). Assessing robustness of intrinsic tests of independence in twoway contingency tables. Tech. rep., Department of Statistics, University of Florida.

Consonni, G. & La Rocca, L. (2008). Tests based on intrinsic priors for the equality of two correlated proportions. *Journal of the American Statistical Association (to appear)* .

Consonni, G. & Massam, H. (2007). Alternative parametrizations and reference priors for decomposable discrete graphical models. Manuscript. arXiv:0707.3873.

Consonni, G. & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science (to appear)* .

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. (1996). *Probabilistic networks and expert systems*. Springer-Verlag, New York.

Dawid, A. P. & Lauritzen, S. L. (2001). Compatible prior distributions. In E. George, ed., *Bayesian methods with applications to science, policy and official statistics*. Monographs of Official Statisticss, Luxembourg, pp. 109–118.

Dellaportas, P. & Forster, J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.

Geiger, D. & Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local independence. *The Annals of Statistics* **25**, 1344–1369.

George, E. (2005). Bayesian model selection. In C. Balakrishnan, N. Read & B. Vidakovic, eds., *Encyclopedia of statistical science*, vol. 1. pp. 418–425.

Knuiman, M. W. & Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**, 1061–1071.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.

Neal, R. (2001). Transferring prior informstion between models using imaginary data. Tech. rep., Department of Statistics, University of Toronto, N. 0108.

Pérez, J. M. & Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* **89**, 491–512.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and bayes factors. In D. Dey & R. C. R., eds., *Handbook of statistics*, vol. 25. Elsevier, Amsterdam, pp. 115–149.

Raftery, A. (1996). Approximate Bayes factor and accounting for model uncertainity in generalized linear models. *Biometrika* **83**, 251–266.

Robert, C. (2001). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer-Verlag, New York.