Journal of Multivariate Analysis 105 (2012) 380–396

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Parametrizations and reference priors for multinomial decomposable graphical models

Guido Consonni [a,*], Hélène Massam [b]

[a] Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia, Via San Felice 5, 27100 Pavia, Italy

[b] Department of Mathematics and Statistics, York University, Toronto, M3J 1P3, Canada

ARTICLE INFO

Article history:
Received 6 October 2010
Available online 12 August 2011

AMS subject classification:
62H17
62F15
62F10

Keywords:
Clique
Conjugate family
Contingency table
Cut
Log-linear model
Multinomial model
Natural exponential family
Reference prior

ABSTRACT

Given a multinomial decomposable graphical model, we identify several alternative parametrizations; in particular we consider conditional probabilities of clique-residuals given separators, as well as generalized log-odds-ratios. For each such parametrization, we construct the corresponding reference prior for suitable groupings of the parameters. Each one of the reference priors we obtain is conjugate to the likelihood and is proper. Furthermore, all these priors are equivalent, in the sense that they can be deduced from each other by a change of variable. We also derive estimators of cell-probabilities based on the reference prior. Finally, we discuss in detail a parametrization associated to a collection of variables representing a cut for the statistical model, and derive the corresponding reference prior.

1. Introduction

Graphical models, see e.g. [27], are statistical models such that dependences between variables are expressed by means of a graph. The study of graphical models is an established and active area of applied and theoretical research.

In this paper, we consider multinomial graphical models Markov with respect to undirected decomposable graphs. While such models are Markov equivalent to Directed Acyclic Graph (DAG) models without immoralities, they are of interest by themselves since they are also used for the analysis of multi-way contingency tables.

We follow a Bayesian approach which requires a prior distribution on the parameter space. Parameter priors for undirected discrete graphical, or more generally log-linear, models have been considered in [17,29,18,26,19].

Despite the adoption of reasonably simplified models, prior elicitation still represents a major concern even for moderately large graphs, because of the very high number of parameters involved. This naturally suggests to search for default, or objective, priors, requiring a minimal subjective input. However there is now evidence, see e.g. [6], that naive approaches based on flat non-informative priors are largely inadequate in multi-parameter settings. In this context, reference analysis represents one of the most successful general methods to derive default prior distributions. For an informative review, see [11]; while Berger et al. [10] provide a formal development. Although the algorithmic complexity

* Corresponding author.
E-mail addresses: guido.consonni@unipv.it (G. Consonni), massamh@yorku.ca (H. Massam).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.08.001

for the construction of reference priors can be substantial, it is known that suitable re-parametrizations of the model may considerably simplify the task; see for instance [15,13].

We address two specific issues in this paper: identifying alternative parametrizations for a given discrete decomposable graphical model, and constructing the corresponding reference priors. More precisely, in Section 2 we consider several parametrizations: conditional probabilities of clique-residuals given separators, as well as generalized log-odds ratios that arise as canonical parameters of equivalent exponential family representations of the underlying sampling distribution, and explicate their mutual relationships. In Section 3, we provide the expressions for the corresponding reference priors, and discuss their main properties. We also provide the expressions for the Bayesian estimator under quadratic loss, i.e. the posterior expectation of the probability of each cell, and for the Bayesian estimator under a normalized squared loss. In Section 4, we present a parametrization associated to a cut in the graphical model, and derive the corresponding reference prior together with the ensuing estimator.

In the last section, we mention some possible points of discussion. In particular, we remark that our estimators could be investigated from the perspective of risk properties, along the lines pursued in [23,24] for Poisson decomposable graphical models. Finally, technical details for the proof of the relationships between various parametrizations are given in the Appendix.

2. Generalized log-odds-ratios parametrizations

2.1. Preliminaries

Let us recall some basic facts about undirected graphs and graphical models: for further details the reader is referred to [27, ch. 2]. An undirected graph G is a pair (V, E) where V is a finite set of vertices and E the set of edges, an edge being an unordered pair $\{\gamma, \delta\},\ \gamma \in V, \delta \in V, \gamma \neq \delta$. Henceforth the graph G is assumed to be decomposable. For a given ordering $C_1, \ldots, C_k$ of the cliques, we define the following sets

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

called, respectively, the l-th history, l-th separator and l-th residual. For a perfect ordering of the cliques (i.e. if for any $l > 1$ there is an $i < l$ such that $S_l \subseteq C_i$) the $S_l,\ l = 2, \ldots, k$ are minimal separators.

A graphical model, Markov with respect to a given graph G, is a family of probability distributions on $(X_\gamma,\ \gamma \in V)$ such that X_δ is independent of X_γ given $X_{V\setminus\{\delta,\gamma\}}$ whenever $\{\gamma, \delta\}$ is not in E. In this paper we shall focus on contingency tables arising from the classification of N units according to a finite set V of criteria, see [27, Ch. 4]. Each criterion is represented by a variable $X_\gamma, \gamma \in V$, which takes values in a finite set $\mathcal{I}_\gamma$. Let $\mathcal{I} = \times_{\gamma\in V} \mathcal{I}_\gamma$. The cells of the table are the elements

$$i = (i_\gamma,\ \gamma \in V), \quad i \in \mathcal{I}. \tag{2.1}$$

Each of N individuals falls into cell i independently with a probability $p(i)$; we let $p = (p(i),\ i \in \mathcal{I})$, with $\sum_{i\in\mathcal{I}} p(i) = 1$. Furthermore, we write $n(i)$ for the i-th cell-count and $n = (n(i),\ i \in \mathcal{I})$, with $\sum_{i\in\mathcal{I}} n(i) = N$. We consider here the model $\mathcal{M}_G$, which, for a given G and a given integer N, is the set of multinomial $\mathcal{M}(N, p)$ distributions, with $N = \sum_{i\in\mathcal{I}} n(i)$ and $p = (p(i),\ i \in \mathcal{I})$ in the $|\mathcal{I}| - 1$ dimensional simplex, which are Markov with respect to G.

From now on, we adopt the notation "$D \subseteq_0 V$" to mean that D may be the empty set while "$D \subseteq V$" excludes the empty set. Let $\mathcal{E}$ denote the power set of V, excluding the empty set, i.e.

$$\mathcal{E} = \{F \subseteq V, F \neq \emptyset\}.$$

For $D \in \mathcal{E}$, we denote the D-marginal cell and its corresponding count by

$$i_D = (i_\gamma,\ \gamma \in D), \quad \text{and} \quad n(i_D), i_D \in \mathcal{I}_D = \times_{\gamma\in D} \mathcal{I}_\gamma. \tag{2.2}$$

We therefore have $n(i_D) = \sum_{j\in\mathcal{I}|j_D=i_D} n(j) = \sum_{j_{V\setminus D}\in\mathcal{I}_{V\setminus D}} n(i_D, j_{V\setminus D})$. Note that $n(i_\emptyset) = N$. For F, D in $\mathcal{E}$, we use the notation $p^D(i_D)$ and $p^{D|i_F}(i_D)$ to denote the marginal and the conditional probabilities, respectively

$$p^D(i_D) = \sum_{j\in\mathcal{I}|j_D=i_D} p(j) \quad \text{and} \quad p^{D|i_F}(i_D) = \frac{p^{D\cup F}(i_D, i_F)}{p^F(i_F)}.$$

Assuming that "0" indicates one of the levels for each variable, we let i^*_γ denote the"0"-level in $\mathcal{I}_\gamma$, so that

$$i^* = (i^*_\gamma,\ \gamma \in V)$$

denotes the cell with all components equal to 0.

Definition 2.1. For $D \in \mathcal{E}$, we define

$$\mathcal{I}^*_D = \{i_D \mid i_\gamma \neq i^*_\gamma,\ \forall\gamma \in D\}. \tag{2.3}$$

In words, $\mathcal{I}^*_D$ is the set of marginal cells i_D such that none of their components is equal to 0. We set $\mathcal{I}^*_V = \mathcal{I}^*$. For example, if $D = \{a, b, c\}$, a takes the values $\{0, 1, 2, 3\}$, b takes the values $\{0, 1, 2\}$, c takes the values $\{0, 1\}$, then

$$\mathcal{I}^*_D = \{(1, 1, 1), (2, 1, 1), (3, 1, 1), (1, 2, 1), (2, 2, 1), (3, 2, 1)\}.$$

2.2. The parametrization for G decomposable

The standard multinomial probability function is usually written in terms of the cell-probabilities $p = (p(i), i \in \mathcal{I})$ as

$$f(n|p) = \frac{N!}{\prod_{i\in\mathcal{I}} n(i)!} \prod_{i\in\mathcal{I}} p(i)^{n(i)}, \tag{2.4}$$

where the only restriction on the parameters $p(i)$ is $\sum_i p(i) = 1$. As shown in [30, Lemma 2.2], if we use the "baseline-constrained" parametrization

$$\theta(i_D) = \log \prod_{F\subseteq_0 D} p(i_F, i^*_{V\setminus F})^{(-1)^{|D\setminus F|}}, \quad D \subseteq_0 V,\ i_D \in \mathcal{I}^*_D, \tag{2.5}$$

the density $\prod_{i\in\mathcal{I}} p(i)^{n(i)}$, with respect to a suitable dominating measure, can be written in natural exponential family (NEF) form $\exp\{\langle\theta, x\rangle - k(\theta)\}$. Note that for $F = \emptyset$, $p(i_F, i^*_{V\setminus F}) = p(i^*)$ and $\theta(i_\emptyset) = \theta(i^*) = \log p(i^*)$. The parameters $\theta(i^*)$ and $p(i^*)$ are not free but functions of the other θ or p parameters. We also emphasize the fact that while $\theta(i_D)$ is indexed by the marginal cell i_D, its definition requires knowledge of the joint probabilities $p(i)$ in the full table. Using Lemma 2.1 in [30] and the Hammersley–Clifford Theorem, it is immediate to show that the multinomial model is Markov with respect to a given decomposable, non complete, graph G, if and only if for $i_D \in \mathcal{I}^*_D$, $D \subseteq V$

$$\theta(i_D) = 0 \quad \text{whenever } D \text{ is not complete in } G. \tag{2.6}$$

In order to give the NEF-representation of the multinomial family, Markov with respect to a decomposable graph G, we define

$$\mathcal{D} = \{D \in \mathcal{E} \mid D \text{ complete}\} \tag{2.7}$$

and more generally, for any subset $A \subseteq V$ of the vertex set,

$$\mathcal{D}^A = \{D \subseteq A \mid D \text{ is complete}\} \quad \text{and} \quad \mathcal{D}^A_0 = \mathcal{D}^A \cap \emptyset \tag{2.8}$$

so that $\mathcal{D} = \mathcal{D}^V$. We are now going to provide the representation of $\mathcal{M}_G$ in three different parametrizations.

The first parametrization is in terms of the log-linear parameters defined in (2.5) with canonical parameter

$$\theta^{mod} = \theta(\mathcal{D}) = (\theta(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}^*_D) \tag{2.9}$$

and corresponding canonical statistic $n(\mathcal{D}) = (n(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}^*_D)$. It will also be convenient to use the notation

$$k(\theta(\mathcal{D}^A)) = \log\left(1 + \sum_{D\subseteq A} \sum_{i_D\in\mathcal{I}^*_D} \exp \sum_{F\subseteq D} \theta(i_F)\right) \tag{2.10}$$

for the cumulant generating function, and the notation $\langle\theta(\mathcal{D}^A), n(\mathcal{D}^A)\rangle = \sum_{D\subseteq A}\sum_{i_D\in\mathcal{I}^*_D} \theta(i_D)n(i_D)$. for the inner product. We note here that, as indicated by our notation, this inner product and (2.10) depend upon $\mathcal{D}^A$ since by (2.6), $\theta(i_F) = 0$ if F is not complete. From the general result (2.22) in [30] it follows immediately, as a special case, that the NEF representation of $\mathcal{M}_G$ in terms of θ^{mod} is as follows.

Proposition 2.1. *Let G be a decomposable graph. The NEF-representation of the multinomial Markov model in terms of the parametrization θ^{mod} is given by*

$$\exp\{\langle\theta(\mathcal{D}), n(\mathcal{D})\rangle - N\, k(\theta(\mathcal{D}))\}. \tag{2.11}$$

Let us now introduce a second parametrization which is relative to the marginal distribution for C_1 and the conditional distributions for R_l given S_l. For a given perfect ordering $C_1, \ldots, C_k$ of the cliques of G, the Markov property implies (see [27], p. 90)

$$\prod_{i\in\mathcal{I}} p(i)^{n(i)} = \prod_{i\in\mathcal{I}} \left(\frac{\prod_{l=1}^k p^{C_l}(i_{C_l})}{\prod_{l=2}^k p^{S_l}(i_{S_l})}\right)^{n(i)} = \prod_{i\in\mathcal{I}} \left(p^{C_1}(i_{C_1}) \prod_{l=2}^k p^{R_l|i_{S_l}}(i_{R_l})\right)^{n(i)}. \tag{2.12}$$

As a consequence we have

$$\prod_{i\in\mathcal{I}} p(i)^{n(i)} = \prod_{i_{C_1}\in\mathcal{I}_{C_1}} (p^{C_1}(i_{C_1}))^{n(i_{C_1})} \prod_{l=2}^k \prod_{i_{S_l}\in\mathcal{I}_{S_l}} \prod_{i_{R_l}\in\mathcal{I}_{R_l}} (p^{R_l|i_{S_l}}(i_{R_l}))^{n(i_{C_l})}. \tag{2.13}$$

Note that (2.13) expresses the multinomial Markov model in terms of the marginal probabilities in the C_1-table, as well as the conditional probabilities in the i_{S_l}-slice of the R_l-table, for $l = 2, \ldots, k$. Expression (2.13) gives rise to a new parametrization which we label p^{cond}

$$p^{cond} = (p^{C_1}, p^{R_l|i_{S_l}},\ i_{S_l} \in \mathcal{I}_{S_l},\ l = 2, \ldots, k), \tag{2.14}$$

where

$$p^{C_1} = (p^{C_1}(i_{C_1}),\ i_{C_1} \in \mathcal{I}_{C_1}) \quad \text{and} \quad p^{R_l|i_{S_l}} = (p^{R_l|i_{S_l}}(i_{R_l}),\ i_{R_l} \in \mathcal{I}_{R_l}).$$

Note that there are $1 + \sum_{l=2}^{k} |\mathcal{I}_{S_l}|$ groups of parameters. We remark that the factorization (2.13), and the allied parametrization p^{cond}, are predicated on a specific ordering of the cliques $C_1, \ldots, C_k$. It is well known that to each undirected decomposable graphical model there corresponds a directed acyclic graph (DAG) model which is Markov equivalent, see [2]. In general, however, a DAG model need not be Markov equivalent to an undirected graphical model. Yet, each DAG model admits a recursive factorization such as (2.13) for the joint density, and thus a parametrization of type p^{cond} can always be defined also for DAG models. The latter will contain the conditional probabilities of each *child*-node given all its *parent*-configurations.

We now define the log-linear parameters corresponding to the factorization (2.13).

Definition 2.2. For each clique C_l, $l = 1, \ldots, k$, we define

$$\theta^{C_l}(i_D) = \log \prod_{F \subseteq_0 D} (p^{C_l}(i_F, i^*_{C_l \setminus F}))^{(-1)^{|D \setminus F|}}, \quad D \subseteq C_l,\ i_D \in \mathcal{I}^*_D. \tag{2.15}$$

Definition 2.3. For each residual R_l, $l = 2, \ldots, k$, and fixed $i_{S_l} \in \mathcal{I}_{S_l}$, we define

$$\theta^{R_l|i_{S_l}}(i_D) = \log \prod_{F \subseteq_0 D} (p^{R_l|i_{S_l}}(i_F, i^*_{R_l \setminus F}))^{(-1)^{|D \setminus F|}}, \quad D \subseteq R_l,\ i_D \in \mathcal{I}^*_D. \tag{2.16}$$

Note that both $\theta^{C_l}(i_D)$ and $\theta^{R_l|i_{S_l}}(i_D)$ are "marginal" parameters, in the sense that they are functions of probabilities in the C_l-marginal table.

For any $A \subseteq V$, $B \subseteq V$, $B \cap A = \emptyset$ and any fixed $i_B \in \mathcal{I}_B$, we write

$$\theta(\mathcal{D}^{C_1}) = (\theta^{C_1}(i_D),\ D \subseteq C_1,\ i_D \in \mathcal{I}^*_D), \tag{2.17}$$

for the log-linear parameters for the clique-C_1-table and $n(\mathcal{D}^{C_1}) = (n(i_D),\ D \subseteq C_1,\ i_D \in \mathcal{I}^*_D)$ for the cell-counts in that table. Similarly,

$$\theta(i_B, \mathcal{D}^A) = (\theta^{A|i_B}(i_D),\ D \subseteq A,\ i_D \in \mathcal{I}^*_D), \tag{2.18}$$

are the log-linear parameters in the i_B-slice of the A-table and $n(i_B, \mathcal{D}^A) = (n(i_B, i_D),\ D \subseteq A,\ i_D \in \mathcal{I}^*_D)$ the cell-counts in that table. We collect together the elements of (2.17) and (2.18) in a single parameter

$$\theta^{cond} = (\theta(\mathcal{D}^{C_1}),\ \theta(i_{S_l}, \mathcal{D}^{R_l}),\ i_{S_l} \in \mathcal{I}_{S_l}, l = 2, \ldots, k) \tag{2.19}$$

with corresponding canonical statistics $n^{cond} = (n(\mathcal{D}^{C_1}),\ n(i_{S_l}, \mathcal{D}^{R_l}),\ i_{S_l} \in \mathcal{I}_{S_l}, l = 2, \ldots, k)$.

Since C_1 and R_l, $l = 2, \ldots, k$, are complete, we can apply Lemma 2.2 of [30] to each of the C_1-marginal and R_l-conditional multinomials in the i_{S_l}-slice of (2.13) and obtain the following representation.

Lemma 2.1. *The NEF-representation, in terms of the parametrization θ^{cond},*

- *of the marginal C_1-model is given by*

$$\prod_{i_{C_1} \in \mathcal{I}_{C_1}} (p^{C_1}(i_{C_1}))^{n(i_{C_1})} = \exp\{\langle \theta(\mathcal{D}^{C_1}), n(\mathcal{D}^{C_1}) \rangle - N\, k(\theta(\mathcal{D}^{C_1}))\} \tag{2.20}$$

- *of the conditional R_l-model in the i_{S_l}-slice is given by*

$$\prod_{i_{R_l} \in \mathcal{I}_{R_l}} (p^{R_l|i_{S_l}}(i_{R_l}))^{n(i_{C_l})} = \exp\{\langle \theta(i_{S_l}, \mathcal{D}^{R_l}), n(i_{S_l}, \mathcal{D}^{R_l}) \rangle - n(i_{S_l})\, k(\theta(i_{S_l}, \mathcal{D}^{R_l}))\}. \tag{2.21}$$

Note that the number of parameters in θ^{mod} and θ^{cond} is of course the same. Indeed each element of each one of the two parametrizations is indexed by i_D, $D \in \mathcal{D}$, $i_D \in \mathcal{I}^*_D$ either directly as for θ^{mod}, or through the components i_F, $F \subseteq S_l$, $i_F \in \mathcal{I}^*_F$ and i_D, $D \subseteq R_l$, $i_D \in \mathcal{I}^*_D$ as for θ^{cond}. Furthermore, θ^{cond} is a block-wise one-to-one function of p^{cond}.

Since the clique marginal generalized log-odds ratios are also of interest, we are now going to define a third parametrization of the multinomial model in terms of the generalized log-odds ratios in (2.15). Any marginal cell i_{S_l} can be written as

$$i_{S_l} = (i_F, i^*_{S_l \setminus F})$$

where $F \subseteq_0 S_l,\ i_{S_l} \in \mathcal{I}^*_{S_l}$ Accordingly, we define

$$\theta(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}) = (\theta^{C_l}(i_F, i_D),\ D \subseteq R_l,\ i_D \in \mathcal{I}^*_D,\ F \subseteq_0 S_l,\ i_F \in \mathcal{I}^*_F)$$
$$n(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}) = (n(i_F, i_D),\ D \subseteq R_l,\ i_D \in \mathcal{I}^*_D,\ F \subseteq_0 S_l,\ i_F \in \mathcal{I}^*_F)$$

and

$$\theta^{cliq} = (\theta(\mathcal{D}^{C_1}),\ \theta(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}),\ l = 2, \ldots, k) \tag{2.22}$$

with corresponding cell counts $n^{cliq} = (n(\mathcal{D}^{C_1}),\ n(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}),\ l = 2, \ldots, k)$. We note that for $F = \emptyset$, $\theta^{C_l}(i_F, i_D) = \theta^{C_l}(i_D)$ and $n(i_F, i_D) = n(i_D)$. Clearly the number of parameters in θ^{cliq} is the same as in θ^{cond}.

The expression of the density in terms of this new parametrization will be given in the next section, after we have derived the relationship between the three parametrizations (2.9), (2.19) and (2.22).

2.3. Relationship between the various θ parametrizations

The relationship between the three θ parametrizations is given in the following proposition. To state the results succinctly, let us also define, for any $F \subseteq V$ and $i_F \in \mathcal{I}^*_F$,

$$i_{\subseteq_0 F} = \{i_G,\ G \subseteq_0 F\}.$$

Then for given $F \subseteq V$ and $i_F \in \mathcal{I}^*_F$, and $A \subseteq V$ such that $F \cap A = \emptyset$, we define

$$\theta(i_{\subseteq_0 F}, \mathcal{D}^A) = (\theta(i_G, j_L),\ G \subseteq_0 F,\ L \subseteq A,\ j_L \in \mathcal{I}^*_L) \tag{2.23}$$

and

$$k(\theta(i_{\subseteq_0 F}, \mathcal{D}^A)) = \log\left(1 + \sum_{L \subseteq A} \exp \sum_{\substack{K \subseteq_0 F \\ H \subseteq L, \\ j_H \in \mathcal{I}^*_H}} \theta(i_K, j_H)\right). \tag{2.24}$$

We note that for any $l = 2, \ldots, k$, and $F \subseteq S_l$,

$$\theta(i_{\subseteq_0 F}, \mathcal{D}^{R_l}) \subset \theta(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}).$$

Proposition 2.2. *Let $i_D \in \mathcal{I}^*_D$ and $D \subseteq C_l$, $D \cap R_l \neq \emptyset$. Then*

(a) *the relationship between θ^{cliq} and θ^{cond} is*

$$\theta^{C_l}(i_D) = \sum_{F \subseteq_0 D \cap S_l} (-1)^{|(D \cap S_l) \setminus F|}\, \theta^{R_l | (i_F, i^*_{S_l \setminus F})}(i_{D \cap R_l}) \tag{2.25}$$

which, for $D \subseteq R_l$, is equivalent to

$$\theta^{R_l | (i_F, i^*_{S_l \setminus F})}(i_D) = \sum_{G \subseteq_0 F} \theta^{C_l}(i_G, i_D). \tag{2.26}$$

(b) *The relationship between θ^{cliq} and θ^{mod} is as follows. Let $\{>l\}$ denote the set of $j \in \{l+1, \ldots, k\}$ such that $C_l \cap C_j \neq \emptyset$.*

(i) *For $D \not\subseteq S_j$, for some $j \in \{>l\}$,*

$$\theta(i_D) = \theta^{C_l}(i_D). \tag{2.27}$$

(ii) *For $D \subseteq S_m$, $m \in \{>l\}$*

$$\theta(i_D) = \theta^{C_l}(i_D) - \sum_{F \subseteq_0 D} (-1)^{|D \setminus F|} k(\theta(i_{\subseteq_0 F}, \mathcal{D}^{C_{>l}})) \tag{2.28}$$

where $C_{>l} = \cup_{m>l}(C_m \setminus C_l)$ and $k(\theta(i_{\subseteq_0 F}, \mathcal{D}^{C_{>l}}))$ is defined as in (2.24).

Moreover, all $\theta(i_H, j_G) \in \theta(i_{\subseteq_0 F}, \mathcal{D}^{C_{>l}})$ are such that $H \cup G \subseteq C_m$ for some $m \in \{>l\}$ and is therefore either equal to $\theta^{C_m}(i_H, j_G)$ or can be expressed in terms of $\theta^{C_m}(i_E)$, $m \in \{>l\}$, $E \subseteq C_m$, $i_E \in \mathcal{I}^*_E$.

The proofs of (2.25) and (2.26) can easily be derived from Definitions 2.2 and 2.3. The proof of (2.28), though, is not immediate and is interesting. It is given in the Appendix.

Remarks. 1. Expression (2.26) is a generalization of the relationship between conditional and marginal log-odds ratios for a three way table given in [1, p. 322].
2. According to (2.27) and (2.28), $\theta(i_D)$ is a function of $\theta^{C_m}(j_H)$ such that $H \subseteq C_m$ for $m \geq l$ only. This is going to be an important fact when we derive the reference prior of θ^{model} from the reference prior on θ^{cond} in the next section.

   Relations (2.27) and (2.28) are crucial for the derivation of the reference prior for θ^{cliq} in the next section, and we therefore illustrate it here with an example.

Example 2.1. Consider a decomposable graphical model with the following perfect order of the cliques

$$C_1 = \{a, b, c\}, \qquad C_2 = \{b, c, d\}, \qquad C_3 = \{c, d, e\}, \qquad C_4 = \{e, f\},$$

having separators

$$S_2 = \{b, c\}, \qquad S_3 = \{c, d\}, \qquad S_4 = \{e\}.$$

To simplify matters, let us assume the data are binary. In this case we can simplify the notation since, because of the corner constraint conditions (see end of Section 2.2), $\mathcal{I}_D^*$ contains only one element for each D. Thus $\theta(i_D)$ can more simply be written $\theta(D)$. Let us take $D = \{c, d\}$. We see that $D \subseteq C_2$ and $D \cap R_2 = \{d\} \neq \emptyset$. Moreover $C_{>2} = \{e, f\}$ and the set of $L \subseteq C_{>2}$ is equal to $\{e, f, ef\}$. Then according to (2.28), it follows that

$$\begin{aligned}\theta(cd) = {} & \theta^{C_2}(cd) - \log(1 + e^{\theta(e)+\theta(ec)+\theta(ed)+\theta(ecd)} + e^{\theta(f)} + e^{\theta(e)+\theta(ec)+\theta(ed)+\theta(ecd)+\theta(f)+\theta(ef)}) \\ & + \log(1 + e^{\theta(e)+\theta(ed)} + e^{\theta(f)} + e^{\theta(e)+\theta(ed)+\theta(f)+\theta(ef)}) + \log(1 + e^{\theta(e)+\theta(ec)} + e^{\theta(f)} + e^{\theta(e)+\theta(ec)+\theta(f)+\theta(ef)}) \\ & - \log(1 + e^{\theta(e)} + e^{\theta(f)} + e^{\theta(e)+\theta(f)+\theta(ef)}).\end{aligned}$$

Since

$$\theta(ec) = \theta^{C_3}(ec), \qquad \theta(ed) = \theta^{C_3}(ed), \qquad \theta(ecd) = \theta^{C_3}(ecd), \qquad \theta(ef) = \theta^{C_4}(ef), \qquad \theta(f) = \theta^{C_4}(f),$$

and according to (2.28) again,

$$\theta(e) = \theta^{C_3}(e) + \log(1 + e^{\theta(f)}) - \log(1 + e^{\theta(f)+\theta(ef)}) = \theta^{C_3}(e) + \log(1 + e^{\theta^{C_4}(f)}) - \log(1 + e^{\theta^{C_4}(f)+\theta^{C_4}(ef)}),$$

we see that $\theta(cd)$ can be expressed in terms of $\theta^{C_m}(E)$, $m \geq 2$, $E \subseteq C_m$.

We will now give the expression of the multinomial Markov model with respect to θ^{cliq}, using relation (2.26).

Lemma 2.2. *Let G be a decomposable graph with its cliques $C_1, \ldots, C_k$ arranged in a perfect order. The NEF-representation of the multinomial Markov model in terms of the θ^{cliq} parametrization is given by*

$$\begin{aligned}\prod_{i \in \mathcal{I}} p(i)^{n(i)} = {} & \exp\{\langle \theta(\mathcal{D}^{C_1}), n(\mathcal{D}^{C_1}) \rangle - N\, k(\theta(\mathcal{D}^{C_1}))\} \prod_{l=2}^{k} \exp\left\{ \langle \theta(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}), n(\mathcal{D}_0^{S_l}, \mathcal{D}^{R_l}) \rangle \right. \\ & \left. - \sum_{F \subseteq_0 S_l} \sum_{j_F \in \mathcal{I}_F^*} n(j_F) \sum_{H \subseteq_0 F} (-1)^{|F \setminus H|} k(\theta(j_{\subseteq_0 H}, \mathcal{D}^{R_l})) \right\}. \end{aligned} \tag{2.29}$$

From (2.29), it appears that under the multinomial Markov model, the joint distribution of n^{cliq} admits a conditional reducibility structure, see [15]; specifically, it factorizes into the product of k conditional exponential families (save for the first term which is a marginal distribution), in a recursive fashion according to the clique ordering.

3. Reference priors

Subjective prior elicitation to express uncertainty on the parameters described in the previous section presents a formidable task, whose difficulty increases with the vertex size. This motivates the use of *objective* priors, which only require as input the statistical model. Even setting aside the enormous difficulty of prior elicitation, a further reason for an objective analysis is to acquire some sort of *consensus* on the elicited prior for scientific inference; see [7]. Often this consensus prior is understood to represent weak prior information, implying that scientific inference would be most convincing for the relevant community of scientists if the result were dominated by data acquired from an experiment, rather than by prior information. A further reason for performing an objective Bayesian analysis is to provide a *benchmark* relative to which a subjective analysis can be compared. Indeed the term reference prior, which now characterizes a specific methodology for the construction of objective priors, stems from this interpretation.

Notice that naive "flat" priors do not represent sensible noninformative distributions in highly dimensional problems, and indeed they should be strongly discouraged, as argued in [11] and references therein. Instead one should start by providing a meaningful definition of a prior which is maximally dominated by the data; this was conceptualized in the seminal paper of Bernardo [12]. Using information-theoretic ideas, he introduced the notion of *reference prior* for a parameter of interest indexing a given model. This is a prior which, within a class of possible distributions, *maximizes the missing information* on the parameter associated to the prior. The concept of missing information requires computing the expected (Shannon) information on the parameter which can be provided by an experiment under hypothetical independent infinite replications. For a rigorous definition, see [10].

When the model contains many parameters, convincing arguments show that one should not maximize the joint missing information, but rather proceed *sequentially*. Berger and Bernardo [8] provide a precise definition of the sequential process required to construct a reference prior.

An important issue which arises when constructing a reference prior is the *ordering* in which the parameters of the model should be considered, since the resulting reference prior will depend on such ordering. This happens because the construction algorithm sequentially maximizes the missing information for each parameter component in the given order (essentially, at each step, one considers the distribution of the current parameter conditional on those following it in the chosen order). Parameters should be arranged in *decreasing* order of inferential interest, so that the first parameter in the list should represent the primary parameter for the investigation. The fact that the ordering of the parameter is relevant to the prior may look odd when regarded from a purely subjective elicitation viewpoint. On the other hand, as Bernardo [11, p. 48], remarks, "reference priors are not meant to describe the analyst's beliefs, but the mathematical formulation of a particular type of prior beliefs those which would maximize the expected missing information about the quantity of interest which could be adopted by consensus as a standard for scientific communication".

Another more subtle issue concerns *grouping* of the parameters. For instance, if the model parameters are $(\phi_1, \phi_2, \ldots, \phi_m)$ say, then we could create a set of $k \leq m$ *groups* $(\phi_{(1)}, \ldots, \phi_{(k)})$, where $\phi_{(1)} = (\phi_1, \ldots, \phi_{m_1})$, $\phi_{(2)} = (\phi_{m_1+1}, \ldots, \phi_{m_1+m_2}), \ldots, \phi_{(k)} = (\phi_{m_1+m_2+\cdots m_{k-1}+1}, \ldots, \phi_m)$. The reference prior on $(\phi_1, \phi_2, \ldots, \phi_m)$ will then depend also on this particular grouping, as well as on the ordering $(\phi_{(1)}, \ldots, \phi_{(k)})$, of the parameters. For a discussion of ordered group reference priors, see [9].

The preceding discussion shows that *weak information* must be focussed on specific aspects, or goals, of the analysis, and the resulting prior will reflect these inputs. In other words, there is no unambiguous definition of the expression "noninformative prior", valid in absolute term. Ideally, ordering and grouping should have a substantive connection to the structure of the problem under consideration to be compelling for an objective Bayesian analysis. Anticipating results that we will present later in this section, we will derive a reference prior for the parameter of a graphical multinomial model which is specific to a particular grouping of the parameters. The latter is inherent in the nature of the graph (and thus has a structural interpretation); on the other hand, the particular order of the groups will prove to be immaterial.

A final issue concerns *invariance* of the reference prior methodology to model reparametrization. As with other techniques which construct priors using an algorithm taking as input a statistical model, there is no guarantee that the prior obtained in the original parametrization ϕ, say, will be coherent with that produced on a newly defined reparametrization of the original model ψ. To be coherent, the two priors should be probabilistically equivalent, i.e. linked to each other through the usual change-of-variable rule of probability theory. A well known situation where this lack of coherence may arise is with conjugate families of priors; see [14,22]. On the other hand, Jeffreys priors are well known to satisfy the property of invariance, which they inherit from the invariance of the expected Fisher information matrix.

Concerning the reference prior, the following result holds. Let $p^{\phi}(\phi)$ be the reference prior for ϕ, relative to the ordered grouping $(\phi_{(1)}, \ldots, \phi_{(k)})$. Let $\psi = g(\phi)$ be a reparametrization and consider the ordered grouping $\psi = (\psi_{(1)}, \ldots, \psi_{(k)})$ with $\psi_{(l)}$ having the same dimension as $\phi_{(l)}$ and $\psi_{(l)}$ being a function of $(\phi_{(1)}, \ldots, \phi_{(l)})$, for $l = 1, \ldots, k$. We call the map $\phi \mapsto \psi$ block-lower-triangular. Then the reference prior $p^{\psi}(\psi) = p^{\phi}(g^{-1}(\psi))|J_{\phi}(\psi)|$, where $J_{\phi}(\psi)$ is the Jacobian of the transformation, so that invariance is satisfied; see [33,16]. Intuitively this happens because the groups of the new parameter ψ are arranged in an order of importance consistent with that of ϕ. An important special case occurs when $\psi_{(l)} = h_l(\phi_{(l)})$: in this case we say that the map is block-wise one-to-one. Clearly invariance is desirable whenever no "natural" or "privileged" parametrization is available for the problem at hand.

We now consider reference priors for the various parametrizations introduced in Section 2. We shall only provide an outline of the proofs, because they follow the technical steps described in Sections 2 and 4.2.1 of [15].

Recall that a reference prior for a multidimensional parameter depends on the grouping of its components, as well as the ordering of its groups. For a given graph G, let $C_1, \ldots, C_k$ represent a perfect ordering of the cliques. We first consider the reference prior for the parametrization p^{cond}, see (2.14). Next we will consider the reference priors for θ^{cond}, θ^{cliq}, θ^{mod} following a parallel grouping-structure. We shall see that all these reference priors are equivalent, so that invariance holds. In other words, our objective Bayesian analysis will not depend on the specific parametrization we choose to work with.

Proposition 3.1. *The reference prior for p^{cond} relative to the grouping defined in (2.14) is*

$$\pi^{R}_{p^{cond}}(p^{cond}) \propto \left(\prod_{i_{C_1} \in \mathcal{I}_{C_1}} p(i_{C_1}) \right)^{-\frac{1}{2}} \prod_{l=2}^{k} \prod_{i_{S_l} \in \mathcal{I}_{S_l}} \left(\prod_{i_{R_l} \in \mathcal{I}_{R_l}} p^{R_l|i_{S_l}}(i_{R_l}) \right)^{-\frac{1}{2}} . \tag{3.1}$$

We note that the reference prior for p^{cond} is a product of Jeffreys' priors, one for each of the groups of p^{cond}, and is conjugate to the family (2.13).

Proof. In our setting, we simply need to derive the (Fisher) information matrix. From (2.13) it appears that the likelihood function factorizes into the product of terms, each involving exactly one group of p^{cond}; furthermore each term is a saturated multinomial. Accordingly the information matrix is block-diagonal, and the determinant of each block, using standard results, is easily available. Specifically the first one, corresponding to clique C_1, is given by

$$N \left(\prod_{i_{C_1} \in \mathcal{I}_{C_1}} p(i_{C_1}) \right)^{-1}, \tag{3.2}$$

while for the remaining blocks the determinant is

$$E(n(i_{S_l})|p) \left(\prod_{i_{R_l} \in \mathcal{I}_{R_l}} p^{R_l|i_{S_l}}(i_{R_l}) \right)^{-1}, \quad i_{S_l} \in \mathcal{I}_{S_l}, l = 2, \ldots, k. \tag{3.3}$$

Because of the perfect ordering the cliques, $S_l \subseteq C_j$ for some $j < l$, so that the expected value $E(n(i_{S_l})|p)$ is a function of parameters only belonging to groups preceding the l-th one. Following the theory summarized in [15, Sect. 2], the reference prior is given by the square root of the product of the block-determinants, excluding the terms $E(n(i_{S_l})|p)$, and the result is established. □

We now emphasize three properties of the reference prior for p^{cond}. First of all, since the information matrix is block-diagonal, the reference prior is order-invariant, i.e. it does not depend on the order of the groups. On the other hand, we recall that the very structure of the parametrization depends on the specific ordering of the cliques $C_1, \ldots, C_k$. Secondly, we remark that there exists also some degree of invariance with respect to grouping. Specifically, if we lumped together in one single block all the i_{S_l} terms $p^{R_l|i_{S_l}}$, $i_{S_l} \in \mathcal{I}_{S_l}$, the reference prior would not change. This feature will turn out to be useful later on when deriving reference priors for alternative parametrizations. Third, we remark that the distribution $\pi^R_{p^{cond}}$ belongs to a family conjugate to the likelihood for p^{cond}, see (2.13). Accordingly its hyper-parameters can be interpreted in terms of "prior counts"; the latter however cannot be recovered as the margins of a fictitious overall table. Indeed, each cell in the C_1-table, as well as in the i_{S_l} slice of the R_l-table, has a prior count equal to $1/2$, irrespective of the dimension of the subtables and of the overall table. Finally, the prior is proper, since it is a product of Dirichlet priors, one for each block, each Dirichlet being indexed by a vector of hyper-parameters with entries all equal to $1/2$. We notice that the expression of the reference prior (3.1) for the p^{cond} parametrization also holds for the corresponding parametrization in an arbitrary DAG model, which contains the conditional probabilities of each child-node given its parents. Indeed, all steps in the proof go through also for this more general class of models.

We now turn to the derivation of the reference priors for the three θ parametrizations described in Section 2. Central to our arguments below is invariance of the reference prior to block-lower-triangular parameter transformations.

We start by expressing p^{cond} in terms of the θ^{cond}. Using (2.15) and (2.16) we therefore define

$$\xi^{C_1}(i_D) = \sum_{F \subseteq D} \theta^{C_1}(i_F), \quad i_D \in \mathcal{I}^*_D \tag{3.4}$$

$$\xi^{R_l|i_F}(i_F, i_D) = \sum_{L \subseteq D} \theta^{R_l|(i_F, i^*_{S_l \setminus F})}(i_L) \tag{3.5}$$

$$= \sum_{L \subseteq D} \sum_{H \subseteq_0 F} \theta^{C_l}(i_H, i_L), \quad i_F \in \mathcal{I}^*_F, i_L \in \mathcal{I}^*_L, D \subseteq R_l. \tag{3.6}$$

We let

$$\xi^{cond} = (\xi^{C_1}, \ \xi^{R_l|(i_F, i^*_{S_l \setminus F})}, \ F \subseteq S_l, i_F \in \mathcal{I}^*_F, \ l = 2, \ldots, k) \tag{3.7}$$

where

$$\xi^{C_1} = (\xi^{C_1}(i_D), D \subseteq C_1)$$
$$\xi^{R_l|(i_F, i^*_{S_l \setminus F})} = (\xi^{R_l|(i_F, i^*_{S_l \setminus F})}(i_F, i_D), \ D \subseteq R_l, i_D \in \mathcal{I}^*_D).$$

The mapping between p^{cond} and ξ^{cond} is block-wise one-to-one. As a consequence the reference prior on ξ^{cond} can be deduced from that of p^{cond} as

$$\pi^R_{\xi^{cond}}(\xi^{cond}) = \pi^R_{p^{cond}}(p^{cond}(\xi^{cond}))|J_{p^{cond}}(\xi^{cond})|, \tag{3.8}$$

where $|J_{p^{cond}}(\xi^{cond})|$ is the absolute value of the Jacobian of the transformation $p^{cond} \mapsto \xi^{cond}$. It can be verified that

$$\det\left(\frac{dp^{cond}}{d\xi^{cond}}\right) = \left(\prod_{i_{C_1}\in\mathcal{I}_{C_1}} p(i_{C_1})(\xi^{C_1})\right)\prod_{l=2}^{k}\prod_{i_{S_l}\in\mathcal{I}_{S_l}}\left(\prod_{i_{R_l}\in\mathcal{I}_{R_l}} p^{R_l|i_{S_l}}(i_{R_l})(\xi^{R_l|i_{S_l}})\right), \tag{3.9}$$

so that the induced reference prior for ξ^{cond} is

$$\pi^R_{\xi^{cond}}(\xi^{cond}) \propto \left(\prod_{i_{C_1}\in\mathcal{I}_{C_1}} p(i_{C_1})(\xi^{C_1})\right)^{-\frac{1}{2}+1}\prod_{l=2}^{k}\prod_{i_{S_l}\in\mathcal{I}_{S_l}}\left(\prod_{i_{R_l}\in\mathcal{I}_{R_l}} p^{R_l|i_{S_l}}(i_{R_l})(\xi^{R_l|i_{S_l}})\right)^{-\frac{1}{2}+1}. \tag{3.10}$$

Clearly the reference prior for ξ^{cond} is also conjugate. We shall also need the following result which can be easily derived from Definitions 2.2 and 2.3 and Moebius inversion formula.

Lemma 3.1. *For $i_{C_1} = (i_F, i^*_{C_1\setminus F})$,*

$$p^{C_l}(i_{C_l}) = \frac{\exp\xi^{C_l}(i_F)}{1+\sum_{H\subseteq C_1}\sum_{j_H\in\mathcal{I}^*_H}\exp\xi^{C_l}_D(j_H)}. \tag{3.11}$$

For i_{S_l} and $i_{R_l} = (i_G, i^*_{R_l\setminus G})$ given,

$$p^{R_l|i_{S_l}}(i_{R_l}) = \frac{\exp\xi^{R_l|i_{S_l}}(i_G)}{1+\sum_{H\subseteq R_l}\sum_{j_H\in\mathcal{I}^*_H}\exp\xi^{R_l|i_{S_l}}_D(j_H)}. \tag{3.12}$$

As particular cases, we have

$$p^{C_l}(i^*_{C_l}) = \frac{1}{1+\sum_{H\subseteq C_1}\sum_{j_H\in\mathcal{I}^*_H}\exp\xi^{C_l}_D(j_H)}, \tag{3.13}$$

and

$$p^{R_l|i_{S_l}}(i^*_{R_l}) = \frac{1}{1+\sum_{H\subseteq R_l}\sum_{j_H\in\mathcal{I}^*_H}\exp\xi^{R_l|i_{S_l}}_D(j_H)}. \tag{3.14}$$

Since the reference priors of the three θ-parametrizations are structurally equivalent we shall provide the result in a unified statement.

Theorem 3.1. *The reference prior for*

(a) θ^{cond}, *relative to the grouping defined in* (2.19)
(b) θ^{cliq}, *relative to the grouping defined in* (2.22)
(c) θ^{mod}, *relative to the following grouping*

$$\tilde\theta^{C_1} = (\theta(i_D), D\subseteq C_1, i_D\in\mathcal{I}^*_D), \qquad \tilde\theta^{C_l} = (\theta(i_D), D\subseteq C_l, D\cap R_l\neq\emptyset), \quad l=2,\ldots,k \tag{3.15}$$

is proportional to

$$\left(\prod_{i_{C_1}\in\mathcal{I}_{C_1}} p(i_{C_1})(\cdot)\right)^{\frac{1}{2}}\prod_{l=2}^{k}\prod_{i_{S_l}\in\mathcal{I}_{S_l}}\left(\prod_{i_{R_l}\in\mathcal{I}_{R_l}} p^{R_l|i_{S_l}}(i_{R_l})(\cdot)\right)^{\frac{1}{2}}, \tag{3.16}$$

where the probabilities $p(i_{C_1})(\cdot)$ and $p^{R_l|i_{S_l}}(i_{R_l})(\cdot)$ are understood to be expressed in terms of the relevant θ-parametrization, using (3.11)–(3.14) *together with* (i) (3.4)–(3.5) *for θ^{cond};* (ii) (3.6) *for θ^{cliq}.* (iii) (2.26)–(2.28) *for θ^{mod}.*

More explicitly, the reference prior

- *for* θ^{cond} *is given by the product of* (2.20) *and* (2.21)*, with the understanding that the counts in these formulas are replaced by fictitious prior counts which we write as* $\tilde{n}(i_D)$, $\tilde{N}$ *and so on. More precisely, we have*

$$\tilde{n}(i_D) = \frac{|\mathcal{I}_{C_1 \setminus D}|}{2}, \qquad \tilde{N} = \frac{|\mathcal{I}_{C_1}|}{2},$$

and

$$\tilde{n}(i_{S_l}, i_D) = \frac{|\mathcal{I}_{R_l \setminus D}|}{2}, \qquad \tilde{n}(i_{S_l}) = \frac{|\mathcal{I}_{R_l}|}{2}.$$

- *for* θ^{cliq} *is given by* (2.29) *where for* $l = 1$

$$\tilde{n}(i_D) = \frac{|\mathcal{I}_{C_1 \setminus D}|}{2} \quad \text{and} \quad \tilde{N} = \frac{|\mathcal{I}_{C_1}|}{2},$$

and for $l = 2, \ldots, k$

$$\tilde{n}(j_F, i_D) = \frac{|\mathcal{I}_{S_l \setminus F}||\mathcal{I}_{R_l \setminus D}|}{2} \quad \text{and} \quad \tilde{n}(j_F) = \frac{|\mathcal{I}_{S_l \setminus F}||\mathcal{I}_{R_l}|}{2}$$

- *for* θ^{mod} *can be obtained from that of* θ^{cliq} *above by expressing it in terms of* $\theta(\mathcal{D})$ *using* (2.27) *and* (2.28)*.*

Proof. (a) Because of (3.4) and (3.5) it is immediate to verify that the map $\xi^{cond} \mapsto \theta^{cond}$ is block-wise one-to-one; moreover the Jacobian is equal to one. Accordingly the reference prior for θ^{cond} will be exactly as that for ξ, with the only difference that the probabilities involved will be expressed as functions of θ^{cond}.

(b) Similarly to what happened for the reference prior for p^{cond}, the reference prior for θ^{cond} is unchanged if, for each $l = 2, \ldots, k$, we lump together the groups labeled by $i_{S_l} \in \mathcal{I}_{S_l}$, and thus only regard θ^{cond} as made up of k groups. In this way the transformation from θ^{cond} to θ^{cliq} is block-wise one-to-one, and thus the reference prior for θ^{cliq} is equal to that induced from the reference prior θ^{cond}. Moreover, the transformation is linear so that the Jacobian is constant, and thus the result follows.

(c) We see that the groupings in (3.15) are exactly parallel to those in θ^{cliq}. From (2.27) and (2.28) we also see that the l-th group in θ^{mod} is a function of the subsequent $l, l+1, \ldots, k$ groups in θ^{cliq}. This defines a block-upper triangular transformation, which can be turned into a block-lower triangular one by reversing the order of the groups in θ^{cliq}. Since the reference prior on θ^{cliq} is invariant to group-ordering, we conclude that the reference prior on θ^{mod} can be obtained from that of θ^{cliq} by a change-of-variable. From (2.27) and (2.28) the Jacobian matrix is upper triangular with diagonal elements $\partial\theta^{C_l}(i_D)/\partial\theta(i_D) = 1$; as a consequence the Jacobian is 1, as one can verify in Example 2.1. Finally, the expressions of the fictitious counts are derived by inspection. □

We remark that, similarly to what happened for p^{cond}, the reference prior for each of the three θ-parametrizations is also a conjugate prior, since each is proportional to the corresponding likelihood, and is proper, being the transformation of a proper prior on p^{cond}.

We now derive the Bayes estimator of the cell probabilities $p(j), j \in \mathcal{I}$ under the reference prior (3.1) and assuming a standard quadratic loss function. This is given by the posterior expectation

$$\begin{aligned} E(p(j) \mid n(i), i \in \mathcal{I}) &= \frac{n(j_{C_1}) + \frac{1}{2}}{\sum_{i_{C_1} \in \mathcal{I}_{C_1}} \left(n(i_{C_1}) + \frac{1}{2}\right)} \prod_{l=2}^{k} \frac{n(j_{C_l}) + \frac{1}{2}}{\sum_{i_{R_l} \in \mathcal{I}_{R_l}} \left(n(j_{S_l}, i_{R_l}) + \frac{1}{2}\right)} \\ &= \frac{n(j_{C_1}) + \frac{1}{2}}{N + \frac{|\mathcal{I}_{C_1}|}{2}} \prod_{l=2}^{k} \frac{n(j_{C_l}) + \frac{1}{2}}{n(j_{S_l}) + \frac{|\mathcal{I}_{R_l}|}{2}} \\ &= \hat{p}(j) \times \left(\frac{1 + \frac{1}{2n(j_{C_1})}}{1 + \frac{|\mathcal{I}_{C_1}|}{2N}} \prod_{l=2}^{k} \frac{1 + \frac{1}{2n(j_{C_l})}}{1 + \frac{|\mathcal{I}_{R_l}|}{2n(j_{S_l})}} \right) \end{aligned} \tag{3.17}$$

where

$$\hat{p}(j) = \left(\frac{n(j_{C_1})}{N} \prod_{l=2}^{k} \frac{n(j_{C_l})}{n(j_{S_l})} \right)$$

is the maximum likelihood estimator of $p(j)$. We recall that the admissibility of the MLE in decomposable log-linear interaction models for multinomial contingency tables was proved in [31] when the loss function is the sum of the squared error losses for each component. Formula (3.17) is easily obtained using the likelihood (2.4) and the reference prior (3.1) and reveals that this estimator is always well defined; in particular it is never zero, even for sparse tables wherein some cliques may present no cases, as opposed to $\hat{p}(j)$.

To better appreciate the nature of estimator (3.17), consider the following hypothetical situation. For a given set observations assume that, within each clique, each configuration j_{C_l} is equally frequent. Then $n(i_{C_l}) = A_l, \quad \forall i_{C_l} \in \mathcal{I}_{C_l}$. On the other hand, $\sum_{i_{R_l} \in \mathcal{I}_{R_l}} n(j_{S_l}, i_{R_l}) = n(j_{S_l})$, whence $A_l = \frac{n(j_{S_l})}{|\mathcal{I}_{R_l}|}$. Notice that for $l = 1$ we get $A_1 = \frac{n(j_\emptyset)}{|\mathcal{I}_{C_1}|} = \frac{N}{|\mathcal{I}_{C_1}|}$. Substituting these values into correction factor in (3.17) we get exactly one. Summing up, deviations of our estimator from the MLE will be noticeable for sparse tables and when counts within cliques are not of equal size.

It is also interesting to compute the Bayes estimator of p^{cond} under the same reference prior (3.1) and the normalized squared error loss

$$L(d, p) = \sum_{i \in \mathcal{I}} \frac{(d(i) - p(i))^2}{p(i)}. \tag{3.18}$$

This loss was also considered by Olkin and Sobel [32] to prove admissibility and minimaxity of the MLE for the saturated multinomial model, by Johnstone [25] in the context of Poisson models and more recently by Hara and Takemura [23,24] for decomposable Poisson models. Some straightforward computations lead to the estimator

$$\widetilde{p(j)} = \frac{n(j_{C_1}) - \frac{1}{2}}{N + \frac{|\mathcal{I}_{C_1}|}{2} - 1} \prod_{l=1}^{k} \frac{n(j_{C_l}) - \frac{1}{2}}{n(j_{S_l}) + \frac{|\mathcal{I}_{R_l}|}{2} - 1} \tag{3.19}$$

with $\widetilde{p(j)} = 0$ when $n(j) = 0$. Notice that estimator (3.17) and (3.19) will be similar when cell counts are moderately large; on the other hand, for sparse tables characterized by several zero cell counts, the difference might be appreciable.

4. Parametrizations and reference priors associated to a cut

The reference priors obtained in the previous section were based on a grouping of the parameters defined by the structure of the graph, essentially through a perfect ordering of the cliques (and consequently of residuals and separators).

Now suppose we are interested in a particular subset $A \subseteq V$ of the variables, and that we would like to consider a reference prior where the first group contain exclusively the parameters of the marginal distribution for the variables in A. As an example, consider the Czech Autoworkers data presented in [20]. They involve 1841 men cross-classified according to 6 binary variables, each representing a potential risk factor for coronary trombosis: (a) smoking, (b) strenuous mental work, (c) strenuous physical work, (d) systolic blood pressure, (e) ratio of beta and alpha lipoproteins and (f) family anamnesis of coronary heart disease. These data were analyzed from a Bayesian model search perspective in [28,18] and more recently [30]. Each paper adopted a specific methodology, and often several analyzes were performed modifying tuning parameters in order to assess robustness. When the space of models was restricted to decomposable graphs, the graph having cliques $C_1 = \{a, d, e\}$, $C_2 = \{a, c, e\}$, $C_3 = \{b, c\}$ and $C_4 = \{f\}$ often emerged as the most probable one. Conditionally on this graph, a goal of the analysis may be to focus on the effect of strenuous physical work on coronary trombosis. Accordingly the parameter associated to the set $A = \{c\}$ would be of primary concern.

In this section we show that, if the Markov model $\mathcal{M}_G$ is *collapsible* onto A (equivalently if A represents a *cut* for the joint distribution), then we are able to obtain the required reference prior.

Asmussen and Edwards [3] consider the concept of collapsibility for contingency tables. If the set of factors for the table are indexed by $\gamma \in V$ and if $A \subseteq V$, we say that G is collapsible onto A if the multinomial model $\mathcal{M}_{G_A}$, Markov with respect to the induced subgraph G_A, is the same as the model obtained by marginalizing the given model $\mathcal{M}_G$, Markov with respect to G, over the A-table. Frydenberg (1990, Theorem 5.4) has shown that the model for the random vector Y, Markov with respect to G, is collapsible onto A if and only if the sub-vector Y_A is a cut (for simplicity we shall also say that A is a cut). Cuts in exponential families have been introduced in [4] and studied in several further articles such as [5]. A very useful result, due to [3], is that A will induce a cut if and only if every connected component of $V \setminus A$ has a complete boundary in G.

Let B_l, $l = 1, \ldots, p$ be the connected components of $V \setminus A$ and let ∂B_l denote the boundary of B_l. The following lemma gives the factorization of $\mathcal{M}_G$ with respect to the cut A and the connected components of $G_{V \setminus A}$.

Lemma 4.1. *Let A be a cut. Let $C'_1, \ldots, C'_q$ be a perfect ordering of the cliques of G_A, the graph induced by A. Let B_l, $l = 1, \ldots, p$ be the connected components of $G_{V \setminus A}$. Let $C_j^{(l)}$, $j = 1, \ldots, m_l$ be the cliques of the induced graph $G_{B_l \cup \partial B_l}$, $l = 1, \ldots, p$. The multinomial model $\mathcal{M}_G$, Markov with respect to G, can be factorized as follows*

$$\prod_{i \in \mathcal{I}} p(i)^{n(i)} = \prod_{i_{C'_1} \in \mathcal{I}_{C'_1}} (p^{C'_1}(i_{C'_1}))^{n(i_{C'_1})} \prod_{l=2}^{q} \prod_{i_{S'_l} \in \mathcal{I}_{S'_l}} \prod_{i_{R'_l} \in \mathcal{I}_{R'_l}} (p^{R'_l | i_{S'_l}}(i_{R'_l}))^{n(i_{C'_l})}$$

$$\times \prod_{l=1}^{p} \left(\prod_{i_{\partial B_l}} \prod_{i_{C_1^{(l)} \setminus \partial B_l}} (p^{C_1^{(l)} \setminus \partial B_l | i_{\partial B_l}}(i_{C_1^{(l)} \setminus \partial B_l}))^{n(i_{C_1^{(l)}})} \prod_{j=2}^{m_l} \prod_{i_{S_j^{(l)}}} \prod_{i_{R_j^{(l)}}} (p^{R_j^{(l)} | i_{S_j^{(l)}}}(i_{R_j^{(l)}}))^{n(i_{C_j^{(l)}})} \right). \tag{4.1}$$

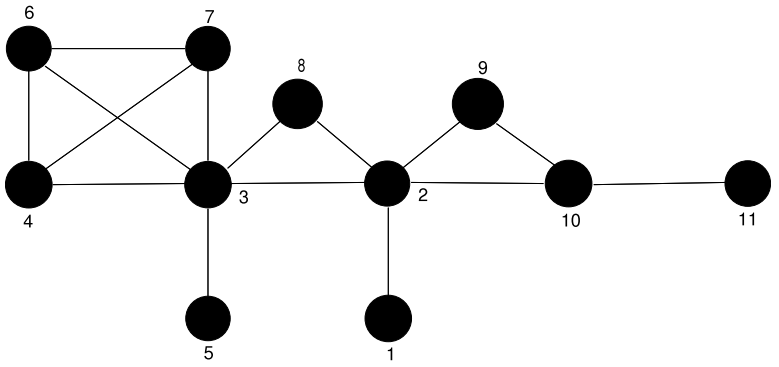

Fig. 1. The decomposable graph for Example 4.1.

Proof. For simplicity of exposition, some statements concerning the random variables associated to a set, will be simply stated in terms of the set itself. If A is a cut, A separates the connected components of $V\setminus A$; by Theorem 2.8 of [17], this implies that the B_l's are mutually conditionally independent given A. Moreover since A is a cut, the boundary of B_l is a complete subset of A and, of course, it separates B_l from $V\setminus(B_l\cup\partial B_l)$. Therefore the overall multinomial Markov model factorizes as the product of the A-marginal multinomial model, Markov with respect to $\mathcal{M}_{G_A}$, and the product of the conditional multinomial distributions of the B_l's given $i_{\partial B_l}$, $l=1,\ldots,p$,. Since the marginal model for A is Markov with respect to the graph G_A, it factorizes according to a perfect order of the cliques of G_A, in parallel to what was done in Section 3: this proves the first line of (4.1).

Let us now consider the expression for the second line of (4.1). As recalled above, this is given by the product of the conditional multinomial models for B_l, $l=1,\ldots,p$ given $i_{\partial B_l}$. For any $l\in\{1,\ldots,p\}$, as a subgraph of G, the induced graph $G_{B_l\cup\partial B_l}$ is decomposable. Moreover the marginal model for $B_l\cup\partial B_l$ is Markov w.r.t. $G_{B_l\cup\partial B_l}$. This happens because $B_l\cup\partial B_l$ is itself a cut, since the boundary of each connected component of $G_{V\setminus(B_l\cup\partial B_l)}$ clearly belongs to ∂B_l which is complete. Therefore the marginal distribution $\mathcal{M}_{G_{B_l\cup\partial B_l}}$ factorizes according to a perfect order of the cliques of $G_{B_l\cup\partial B_l}$. Since ∂B_l is complete, it must belong to a clique $C_1^{(l)}$ of $G_{B_l\cup\partial B_l}$ and by Proposition 2.29 of [27], we know that we can take this clique as the first in a perfect order $C_i^{(l)}$, $i=1,\ldots,m_l$ of the cliques of $G_{B_l\cup\partial B_l}$.

The marginal multinomial distribution $\mathcal{M}_{G_{B_l\cup\partial B_l}}$ can therefore be written as

$$\begin{aligned}\prod_{j=1}^{m_l}\prod_{i_{C_j^{(l)}}}(p^{C_j^{(l)}}(i_{C_j^{(l)}}))^{n(i_{C_j^{(l)}})} &= \prod_{i_{C_1^{(l)}}} p^{C_1^{(l)}}(i_{C_1^{(l)}})^{n(i_{C_1^{(l)}})}\prod_{j=2}^{m_l}\prod_{i_{S_j^{(l)}}}\prod_{i_{R_j^{(l)}}}(p^{R_j^{(l)}|i_{S_j^{(l)}}}(i_{R_j^{(l)}}))^{n(i_{C_j^{(l)}})}\\ &= \prod_{i_{\partial B_l}}\left((p^{\partial B_l}(i_{\partial B_l}))^{n(i_{\partial B_l})}\prod_{i_{C_1^{(l)}\setminus\partial B_l}}(p^{C_1^{(l)}\setminus\partial B_l|i_{\partial B_l}}(i_{C_1^{(l)}\setminus\partial B_l}))^{n(i_{C_j^{(l)}})}\right)\\ &\quad\times\prod_{j=2}^{m_l}\prod_{i_{S_j^{(l)}}}\prod_{i_{R_j^{(l)}}}p^{R_j^{(l)}|i_{S_j^{(l)}}}(i_{R_j^{(l)}})^{n(i_{C_j^{(l)}})}\end{aligned}$$

and therefore the model for B_l conditional on $i_{\partial B_l}$ is equal to

$$\prod_{i_{C_1^{(l)}\setminus\partial B_l}}(p^{C_1^{(l)}\setminus\partial B_l|i_{\partial B_l}}(i_{C_1^{(l)}\setminus\partial B_l}))^{n(i_{C_j^{(l)}})}\prod_{j=2}^{m_l}\prod_{i_{R_j^{(l)}}}\prod_{i_{S_j^{(l)}}}p^{R_j^{(l)}|i_{S_j^{(l)}}}(i_{R_j^{(l)}})^{n(i_{C_j^{(l)}})}. \tag{4.2}$$

Since this is true for all B_l, $l=1,\ldots,p$, the result is established. □

With reference to the Czech Autoworkers case, in which we had singled out variable c (strenuous physical work) as worth of being investigated on its own, it can be easily checked that c is a cut relative to the most probable model recalled above. On the other hand, if we select variable b (strenuous mental work), then our result does not hold because b is not a cut. A more elaborate artificial example follows in order to better clarify the main points.

Example 4.1. Suppose that the joint distribution of the 11 variables numbered consecutively from 1 to 11 is Markov with respect to the decomposable graph G as given in Fig. 1.

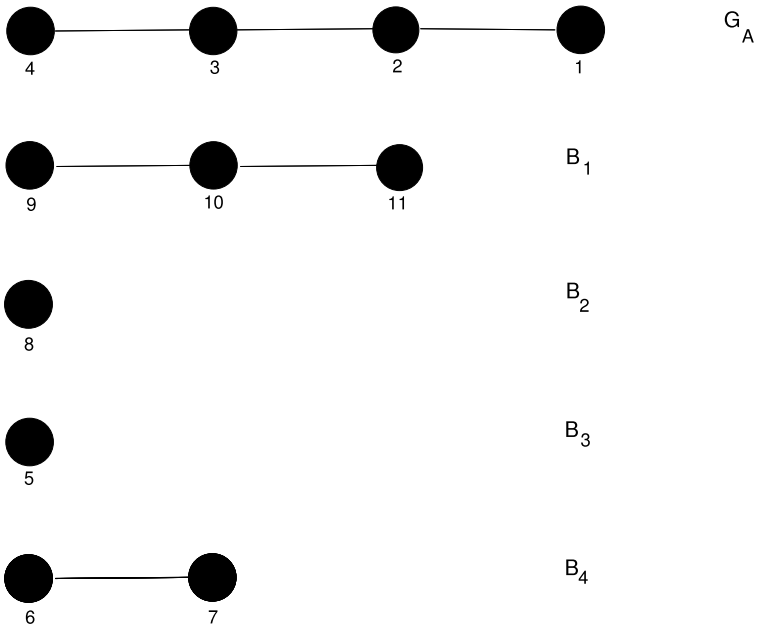

Fig. 2. The decomposable graph G_A associated to a cut A and the connected components of $G_{V\setminus A}$ for Example 4.1.

Consider the subset of variables given by $A = \{1, 2, 3, 4\}$. A perfect ordering of the cliques of the induced sub-graph G_A is

$$C'_1 = \{1, 2\}, \qquad C'_2 = \{2, 3\}, \qquad C'_3 = \{3, 4\}, \tag{4.3}$$

so that $S'_2 = \{2\},\ S'_3 = \{3\}, R'_2 = \{3\},\ R'_3 = \{4\}$. The connected components B_l of $G_{V\setminus A}$, their boundary ∂B_l together with the cliques C_j^l of $G_{B_l\cup\partial B_l}$ are

l	B_l	∂B_l	$B_l \cup \partial B_l$	$C_j^{(l)}$
1	$\{9, 10, 11\}$	$\{2\}$	$\{2, 9, 10, 11\}$	$C_1^{(1)} = \{2, 9, 10\},\ C_2^{(1)} = \{10, 11\}$
2	$\{8\}$	$\{2, 3\}$	$\{2, 3, 8\}$	$C_1^{(2)} = \{2, 3, 8\}$
3	$\{5\}$	$\{3\}$	$\{3, 5\}$	$C_1^{(3)} = \{3, 5\}$
4	$\{6, 7\}$	$\{3, 4\}$	$\{3, 4, 6, 7\}$	$C_1^{(4)} = \{3, 4, 6, 7\}$

A graphical display of G_A and its connected components is given in Fig. 2.
Accordingly, the multinomial model, Markov with respect to G, can be factorized using Lemma 4.1 as

$$\begin{aligned}
\prod_{i\in\mathcal{I}} p(i)^{n(i)} = {} & \prod_{i_{C'_1}\in\mathcal{I}_{C'_1}} (p^{C'_1}(i_{C'_1}))^{n(i_{C'_1})} \prod_{i_{S'_2}\in\mathcal{I}_{S'_2}} \prod_{i_{R'_2}\in\mathcal{I}_{R'_2}} (p^{R'_2|i_{S'_2}}(i_{R'_2}))^{n(i_{C'_2})} \prod_{i_{S'_3}\in\mathcal{I}_{S'_3}} \prod_{i_{R'_3}\in\mathcal{I}_{R'_3}} (p^{R'_3|i_{S'_3}}(i_{R'_3}))^{n(i_{C'_3})} \\
& \times \prod_{i_{\partial B_1}} \prod_{i_{C_1^{(1)}\setminus\partial B_1}} (p^{C_1^{(1)}\setminus\partial B_1|i_{\partial B_1}}(i_{C_1^{(1)}\setminus\partial B_1}))^{n(i_{C_1^{(1)}})} \prod_{i_{S_2^{(1)}}} \prod_{i_{R_2^{(1)}}} (p^{R_2^{(1)}|i_{S_2^{(1)}}}(i_{R_2^{(1)}}))^{n(i_{C_2^{(1)}})} \\
& \times \prod_{i_{\partial B_2}} \prod_{i_{C_1^{(2)}\setminus\partial B_2}} (p^{C_1^{(2)}\setminus\partial B_2|i_{\partial B_2}}(i_{C_1^{(2)}\setminus\partial B_2}))^{n(i_{C_1^{(2)}})} \\
& \times \prod_{i_{\partial B_3}} \prod_{i_{C_1^{(3)}\setminus\partial B_3}} (p^{C_1^{(3)}\setminus\partial B_3|i_{\partial B_3}}(i_{C_1^{(3)}\setminus\partial B_3}))^{n(i_{C_1^{(3)}})} \\
& \times \prod_{i_{\partial B_4}} \prod_{i_{C_1^{(4)}\setminus\partial B_4}} (p^{C_1^{(4)}\setminus\partial B_4|i_{\partial B_4}}(i_{C_1^{(4)}\setminus\partial B_4}))^{n(i_{C_1^{(4)}})}.
\end{aligned}$$

We now provide the expression for the reference prior associated to a cut.

Theorem 4.1. *Let A be a cut and consider the parametrization associated to A*

$$p_A^{cut} = (p_A'^{,\,cond}, p_{V\setminus A|A}'^{,\,cond}),$$

where

$$p_A'^{,\,cond} = (p^{C_1'}, p^{R_l'|i_{S_l'}}, l = 2, \ldots, q,\ i_{S_l'} \in \mathcal{I}_{S_l'}) \tag{4.4}$$

$$p_{V\setminus A|A}'^{,\,cond} = (p^{C_1^{(l)}\setminus \partial B_l | i_{\partial B_l}}, i_{\partial B_l} \in \mathcal{I}_{\partial B_l}; p^{R_j^{(l)}|i_{S_j^{(l)}}}\ l = 1, \ldots, p,\ j = 2, \ldots, m_l,\ i_{S_j^{(l)}} \in \mathcal{I}_{S_j^{(l)}}), \tag{4.5}$$

using the notation presented in Lemma 4.1. The reference prior for p_A^{cut}, relative to the grouping (4.4) and (4.5), is

$$\pi^R_{p_A^{cut}}(p_A^{cut}) \propto \prod_{i_{C_1'} \in \mathcal{I}_{C_1'}} p^{C_1'}(i_{C_1'})^{-\frac{1}{2}} \prod_{l=2}^{q} \prod_{i_{S_l'} \in \mathcal{I}_{S_l'}} \prod_{i_{R_l'} \in \mathcal{I}_{R_l'}} (p^{R_l'|i_{S_l'}}(i_{R_l'}))^{-\frac{1}{2}}$$
$$\times \prod_{l=1}^{p} \left(\prod_{i_{\partial B_l}} \prod_{i_{C_1^{(l)}\setminus \partial B_l}} (p^{C_1^{(l)}\setminus \partial B_l | i_{\partial B_l}}(i_{C_1^{(l)}\setminus \partial B_l}))^{-\frac{1}{2}} \prod_{j=2}^{m_l} \prod_{i_{S_j^{(l)}}} \prod_{i_{R_j^{(l)}}} (p^{R_j^{(l)}|i_{S_j^{(l)}}}(i_{R_j^{(l)}}))^{-\frac{1}{2}} \right).$$

We emphasize that, also for this case, the prior admits a conjugate structure and is proper, being a product of Jeffreys' priors.

Proof. Using Lemma 4.1 the likelihood factorizes into a product of two general terms, one related to the marginal distribution of A indexed by $p_A'^{,\,cond}$, the other related to the conditional distribution of $V\setminus A$ given A indexed by $p_{V\setminus A|A}'^{,\,cond}$. The two groups of parameters are variation and likelihood independent, so that the information matrix is two-block-diagonal. The marginal distribution related to A is a G_A-Markov model, with G_A decomposable, and therefore the corresponding reference prior is exactly as in the general decomposable case of Proposition 3.1. This yields the first line of the kernel of the reference prior

To prove the second line, we have to consider the second block of the information matrix. This actually further decomposes into p diagonal blocks, one for each connected component B_l. Consider the block corresponding to the model for B_l conditional on ∂B_l, $l = 1, \ldots, p$ (see (4.2)). Each block decomposes into $|\mathcal{I}_{\partial B_l}| \times \prod_{j=2}^{m_l} |\mathcal{I}_{S_j^{(l)}}|$ sub-blocks, each one representing the information of a saturated multinomial. In particular the first sub-block has cell-probabilities $p^{C_1^{(l)}\setminus \partial B_l | i_{\partial B_l}}$ and $n(i_{\partial B_l})$ trials, while the remaining sub-blocks have cell-probabilities $p^{R_j^{(l)}|i_{S_j^{(l)}}}$ and $n(i_{S_j^{(l)}})$ trials. The expression of the corresponding term in the information matrix will therefore be as in the general conditional saturated multinomial, see (3.3). Finally, the expectation of $n(i_{\partial B_l})$ depends only on the parameter $p_A'^{,\,cond}$ since $\partial B_l \in A$, and similarly the expectation of $n(i_{S_j^{(l)}})$ does not depend on the parameter $p^{R_j^{(l)}|i_{S_j^{(l)}}}$ specific to the sub-block because of the perfect ordering of the cliques. Therefore, in both cases the term corresponding to the expectation factors out of the determinant and the proof is complete. □

We can, of course, derive the Bayes estimators of cell probabilities under the reference prior in Theorem 4.1 in a parallel way to what was done for the Bayes estimators under the prior given in Proposition 3.1. For the standard quadratic loss function, the Bayes estimator is

$$\mathbf{E}(p(j)|n(i), i \in \mathcal{I}) = \frac{n(j_{C_1'}) + \frac{1}{2}}{N + \frac{|\mathcal{I}_{C_1'}|}{2}} \prod_{l=2}^{q} \frac{n(j_{C_l'}) + \frac{1}{2}}{n(j_{S_l'}) + \frac{|\mathcal{I}_{R_l'}|}{2}} \prod_{l=1}^{p} \frac{n(j_{C_1^{(l)}}) + \frac{1}{2}}{n(j_{\partial B_l}) + \frac{|\mathcal{I}_{C_1^{(l)}\setminus \partial B_l}|}{2}} \prod_{k=2}^{m_l} \frac{n(j_{C_k^{(l)}}) + \frac{1}{2}}{n(j_{S_k^{(l)}}) + \frac{|\mathcal{I}_{R_k^{(l)}}|}{2}}$$

with a similar expression for the Bayes estimator $\widetilde{p(j)}$ under the normalized square error loss.

5. Discussion

In this paper we have considered several alternative parametrizations for discrete decomposable graphical models. Arguing that prior elicitation becomes rapidly infeasible for large graphs, we have adopted an objective Bayes approach analysis which requires no prior input form the user. We have then derived a reference prior for each of the above parametrizations, showing that they are all equivalent. This is reassuring and lends stability to our analysis, which is thus not tied to any particular model-parametrization. A notable feature is that all reference priors are proper and that they belong to a conjugate family; the latter property extends results of [15], valid for Natural Exponential Families having a simple quadratic variance, to multinomial decomposable models, whose variance function is not quadratic.

While our results refer to undirected graphs, they can be of interest for general DAG models with no hidden variables, wherein the joint density exhibits a recursive structure such as (2.13), where the conditional distribution of each single variable given its parents is multinomial, so that the latter admits an exponential family representation. Accordingly both the p^{cond} and θ^{cond} parametrizations can be extended to this larger class of models, together with the corresponding reference priors. On the other hand, since DAG models are generally curved exponential families, see [21], parametrizations θ^{mod} and θ^{clique} are not available for general DAG models.

Using our reference prior, we have derived Bayesian estimators of the cell-probabilities, both under a sum of squared error losses for each component, as well as under a normalized squared error loss. In the former situation, the MLE for the multinomial decomposable case is known to be admissible, as shown by Meeden et al. [31]. Thus a uniform improvement over the risk function associated to the MLE cannot be obtained. This is in contrast to results derived by Hara and Takemura [24,24] for decomposable Poisson graphical models wherein carefully selected parameter priors produce estimators which dominate the MLE for normalized squared error losses. An interesting line of future research would be to compare the latter estimators, in terms of risk behavior, with those originating from an application of our reference prior methodology to the Poisson sampling scheme.

Acknowledgments

We thank two reviewers and the Associate Editor for helpful comments which improved the presentation of this paper. The research of Guido Consonni was supported by MIUR, Italy, PRIN 2007XECZ7L_001, as well as by the University of Pavia. Hélène Massam was supported by NSERC Discovery Grant A8946.

Appendix

Proof of (2.28). Consider $D \subseteq C_l$, $D \cap R_l \neq \emptyset$ for some $l \in \{1, \ldots, k-1\}$ such that also $D \subseteq S_j$ for some $j \in \{>l\}$, then

$$
\begin{aligned}
p^{C_l}(i_D) &= \sum_{L \subseteq_0 C_l^c} p(i_D, j_L) = \sum_{L \subseteq_0 C_l^c} \exp\left(\sum_{E \subseteq_0 D} \theta(i_E) + \sum_{E \subseteq_0 D, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_E, j_G) \right) \\
&= \left(\exp \sum_{E \subseteq_0 D} \theta(i_E) \right) \left(1 + \sum_{L \subseteq C_l^c} \exp\left(\sum_{E \subseteq_0 D, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_E, j_G) \right) \right) \\
\log p^{C_l}(i_D) &= \sum_{E \subseteq_0 D} \theta(i_E) + \log\left(1 + \sum_{L \subseteq C_l^c} \exp\left(\sum_{E \subseteq_0 D, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_E, j_G) \right) \right) \\
\sum_{E \subseteq_0 D} \theta(i_E) &= \log p^{C_l}(i_D) - \log\left(1 + \sum_{L \subseteq C_l^c} \exp\left(\sum_{E \subseteq_0 D, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_E, j_G) \right) \right).
\end{aligned}
$$

This last equality is of the form

$$\sum_{E \subseteq_0 D} \psi(E) = \phi(D) \tag{A.1}$$

and therefore by Moebius inversion formula, we have

$$\psi(D) = \sum_{F \subseteq_0 D} (-1)^{|D \setminus F|} \phi(F). \tag{A.2}$$

For $l = 2, \ldots, k$, let $C_{<l} = H_{l-1} \setminus C_l$. Then (A.2) can be written as

$$
\begin{aligned}
\theta(i_D) &= \sum_{F \subseteq_0 D} (-1)^{|D \setminus F|} \log p^{C_l}(i_F) - \sum_{F \subseteq_0 D} (-1)^{|D \setminus F|} \log\left(1 + \sum_{L \subseteq C_l^c} \exp\left(\sum_{H \subseteq_0 F, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_H, j_G) \right) \right) \\
&= \theta^{C_l}(i_D) - \sum_{F \subseteq_0 D} (-1)^{|D \setminus F|} \log\left(1 + \sum_{L \subseteq C_l^c} \exp\left(\sum_{H \subseteq_0 F, G \subseteq L, j_G \in \mathcal{I}_G^*} \theta(i_H, j_G) \right) \right)
\end{aligned}
$$

$$= \theta^{C_l}(i_D) - \sum_{F \subseteq_0 D} (-1)^{|D\setminus F|} \log \left(1 + \left(\sum_{L \subseteq C_{<l}} + \sum_{L \subseteq C_{>l}} + \sum_{L \subseteq C_{<l} \cup C_{>l}} \right) \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right)$$

$$= \theta^{C_l}(i_D) - \sum_{F \subseteq_0 D} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right) \left(1 + \sum_{L \subseteq C_{>l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right)$$

$$= \theta^{C_l}(i_D) - \sum_{F \subseteq_0 D} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right)$$

$$- \sum_{F \subseteq_0 D} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{>l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right).$$

We now want to show that the term

$$\sum_{F \subseteq_0 D} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right) \tag{A.3}$$

in the equation above is equal to zero.

Let F be an arbitrary subset of D and let $I = F \cap H_{l-1}$. Since $G \subseteq L \subseteq C_{<l}$, in order for $\theta(i_H, j_G)$, $H \subseteq_0 F, j_G \in \mathcal{J}_G^*$ to be non zero, it is necessary that $H \subseteq_0 I$ and therefore

$$\left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right) = \left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 I, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right). \tag{A.4}$$

We see that the right hand side of (A.4) above is the same for all $F \subseteq_0 D$ that have the same intersection I with H_{l-1}. We therefore consider all such F's. Since $D \cap R_l \neq \emptyset$, there are as many such F's with $|D \setminus F|$ odd as there are with $|D \setminus F|$ even and therefore from (A.4), it follows that, for a given I,

$$\sum_{\substack{F \subseteq_0 D \\ F \cap H_{l-1} = I}} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{<l}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right) = 0. \tag{A.5}$$

Since this is true for all $I \subseteq_0 D \cap H_{l-1}$, it follows immediately from (A.5) that (A.3) is equal to zero and we have

$$\theta(i_D) = \theta^{C_l}(i_D) - \sum_{F \subseteq D} (-1)^{|D\setminus F|} \log \left(1 + \sum_{L \subseteq C_{\{>l\}}} \exp \left(\sum_{\substack{H \subseteq_0 F, \\ G \subseteq L, \\ j_G \in \mathcal{J}_G^*}} \theta(i_H, j_G) \right) \right).$$

Formula (2.28) is thus proved. Moreover, since $D \subseteq S_j \cap C_l$ for some $m \in \{>l\}$ and $G \subseteq L \subseteq C_m \setminus C_l$ is non empty, in the right hand side of the equation above, we have that either $\theta(i_H, j_G) = \theta^{C_m}(i_H, j_G)$ or that $\theta(i_H, j_G)$ can be expressed using (2.28) recursively and therefore $\theta(i_D)$ can be expressed in terms of $\theta^{C_m}(i_E)$, $m \in \{>l\}$, $E \subseteq C_m$, $i_E \in \mathcal{J}_E^*$. □

References

[1] A. Agresti, Categorical Data Analysis, 2nd ed., Wiley, Chichetser, 2002.
[2] S.A. Andersson, D. Madigan, M. Perlman, On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs, Scand. J. Stat. 24 (1997) 81–102.
[3] S. Asmussen, D. Edwards, Collapsibility and response variables in contingency tables, Biometrika 70 (1983) 567–578.
[4] O.E. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley, Chichester, 1978.
[5] O.E. Barndorff-Nielsen, A.E. Koudou, Cuts in natural exponential families, Theory Probab. Appl. 40 (1995) 361–372.
[6] J.O. Berger, Bayesian analysis: a look at today and thoughts of tomorrow, J. Amer. Statist. Assoc. 95 (2000) 1269–1276.
[7] J.O. Berger, The case for objective Bayesian analysis (with discussion), Bayesian Anal. 1 (2006) 385–402.
[8] J.O. Berger, J.M. Bernardo, On the development of reference priors, in: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics, vol. 4, Clarendon Press, Oxford, 1992, pp. 35–60.
[9] J.O. Berger, J.M. Bernardo, Ordered group reference priors with application to a multinomial problem, Biometrika 79 (1992) 25–37.
[10] J.O. Berger, J.M. Bernardo, D. Sun, The formal definition of reference priors, Ann. Statist. 37 (2009) 905–938.
[11] J.M. Bernardo, Reference analysis, in: D.K. Dey, C.R. Rao (Eds.), in: Handbook of Statistics, vol. 25, Elsevier, Amsterdam, 2005, pp. 17–90.
[12] J.M. Bernardo, Reference posterior distributions for Bayesian inference, J. R. Statist. Soc. B 41 (1979) 113–147.
[13] G. Consonni, V. Leucari, Reference priors for discrete graphical models, Biometrika 93 (2006) 23–40.
[14] G. Consonni, P. Veronese, Conjugate priors for exponential families having quadratic variance functions, J. Amer. Statist. Assoc. 87 (1995) 1123–1127.
[15] G. Consonni, P. Veronese, Conditionally reducible natural exponential families and enriched conjugate priors, Scand. J. Statist. 28 (2001) 377–406.
[16] G.S. Datta, M. Ghosh, On the invariance of noninformative priors, Ann. Statist. 24 (1996) 141–159.
[17] A.P. Dawid, S.L. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable models, Ann. Statist. 21 (1993) 1272–1317.
[18] P. Dellaportas, J.J. Forster, Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, Biometrika 86 (1999) 615–633.
[19] P. Dellaportas, C. Tarantola, Model determination for categorical data with factor level merging, J. R. Stat. Soc. B 67 (2005) 269–283.
[20] D.E. Edwards, T. Havranek, A fast procedure for model search in multidimensional contingency tables, Biometrika 72 (1985) 339–351.
[21] D. Geiger, D. Heckerman, H. King, C. Meek, Stratified exponential families: graphical models and model selection, Ann. Statist. 29 (2001) 505–529.
[22] E. Gutiérrez-Peña, A.F.M. Smith, Conjugate parameterizations for natural exponential families, J. Amer. Stat. Assoc. 90 (1995) 1347–1356.
[23] H. Hara, A. Takemura, Improving on the maximum likelihood estimators of the means in Poisson decomposable graphical models, J. Multivariate Anal. 98 (2007) 410–434.
[24] H. Hara, A. Takemura, Bayes admissible estimation of the means in Poisson decomposable graphical models, J. Statist. Plann. Inference 139 (2009) 1297–1319.
[25] I. Johnstone, Admissible estimation, Dirichlet principles and recurrence of birth-death chains on Z_+^p, Probab. Theory Related Fields 71 (1986) 231–269.
[26] R. King, S.P. Brooks, Prior induction for log-linear models for general contingency table analysis, Ann. Statist. 29 (2001) 715–747.
[27] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, 1996.
[28] D. Madigan, A. Raftery, Model selection and accounting for model uncertainty in graphical models using Occams window, J. Amer. Statist. Assoc. 89 (1994) 1535–1546.
[29] D. Madigan, J. York, Bayesian graphical models for discrete data, Int. Statist. Rev. 63 (1995) 215–232.
[30] H. Massam, J. Liu, A. Dobra, A conjugate prior for discrete hierarchical log-linear models, Ann. Statist. 37 (2009) 3431–3467.
[31] G. Meeden, C. Geyer, J. Long, E. Funa, The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables, Commun. Stat. — Theory Methods 27 (1998) 473–493.
[32] I. Olkin, M. Sobel, Admissible and minimax estimation for the multinomial distribution and for K independent multinomial distributions, Ann. Statist. 7 (1979) 284–290.
[33] R. Yang, Invariance of the reference prior under reparameterization, Test 4 (1995) 83–94.