Parameterizations and fitting of bi-directed graph models to categorical data

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

Giovanni M. Marchetti
Dipartimento di Statistica "G. Parenti", University of Florence, Italy

Wicher P. Bergsma
London School of Economics and Political Science, London, UK

Abstract

We discuss two parameterizations of models for marginal independencies for discrete distributions which are representable by bi-directed graph models, under the global Markov property. Such models are useful data analytic tools especially if used in combination with other graphical models. The first parameterization, in the saturated case, is also known as the multivariate logistic transformation, the second is a variant that allows, in some (but not all) cases, variation independent parameters. An algorithm for maximum likelihood fitting is proposed, based on an extension of the Aitchison and Silvey method.

Key words: covariance graphs, complete hierarchical parameterizations, connected set Markov property, constrained maximum likelihood, marginal independence, marginal log-linear models, multivariate logistic transformation, variation independence

Running title: Bi-directed graph models for categorical data

1 Introduction

This paper deals with the parametrization and fitting of a class of marginal independence models for multivariate discrete distributions. These models are associated with a class of graphs where the missing edges represent marginal independencies. The graphs used have special edges to distinguish them from undirected graphs used to encode conditional independencies. Cox & Wermuth (1993) use dashed edges and call the graphs covariance graphs by stressing the equivalence between a marginal pairwise independence and a zero covariance in a Gaussian distribution. Richardson & Spirtes (2002) use instead bi-directed edges following the tradition of path analysts. The interpretation of the graphs in terms of independencies is based on the pairwise and global Markov properties discussed originally by Kauermann (1996) for covariance graphs and later developed by Richardson (2003). These are recalled in section 2. Figure 1 provides two examples of bi-directed graph models with the respective independence statements.

Models of marginal independence can be useful in several contexts and sometimes they may be used to represent independence structures induced after marginalizing over latent variables. Cox & Wermuth (1993) present a motivating example on diabetic patients concerning four continuous variables. For further discussion on motivations and the interpretation of bi-directed graph models see Wermuth *et al.* (2006) and recently Drton & Richardson (2008).

[Figure 1 about here.]

Developing a parameterization for Gaussian bi-directed graph models is straightforward since the pairwise and the global Markov property are equivalent and they can be simply fulfilled by constraining a subset of covariances to zero. Accomplishing the same task in the discrete case is much more difficult due to the high number of parameters and to the non-equivalence of the two Markov properties. Recently, Drton & Richardson (2008) studied the parametrization of bi-directed graph models for binary distributions, based on Möebius parameters with a version of their iterative conditional fitting algorithm for maximum likelihood estimation.

In this paper we propose different parameterizations, suitable for general categorical variables, based on the class of marginal log-linear models of Bergsma & Rudas (2002). One special case in this class, useful in the context of bi-directed graph models, is the multivariate logistic parameterization of Glonek & McCullagh (1995); see also Kauermann (1997). We discuss a further marginal log-linear parametrization that can, in some cases, be shown to imply variation independent parameters. We show that the marginal log-linear parameterizations suggest a class of reduced models defined by constraining certain higher-order log-linear parameters to zero. Then we discuss maximum likelihood estimation of the models and propose a general algorithm based on previous works by Aitchison & Silvey (1958), Lang (1996) and Bergsma (1997).

The remainder of this paper is organized as follows. Section 2 reviews discrete bi-directed graphs and their Markov properties. In section 3 we give the essential results concerning the theory of marginal log-linear models. Two parameterizations of bi-directed graph models are then given in section 4 illustrating their properties with special emphasis on variation independence and the interpretation of the parameters. In section 5 we

propose an algorithm for maximum likelihood fitting and then, in section 6 we provide an example based on a data set taken from the U.S. General Social Survey. Finally, in section 7 we give a short discussion, with a comparison with the approach by Drton & Richardson (2008).

2 Discrete bi-directed graph models

Bi-directed graphs are essentially undirected graphs with edges represented by bi-directed arrows instead of full lines. We review in this section the main concepts of graph theory required to understand the models. A bi-directed graph $G = (V, E)$ is a pair (V, E), where $V = \{1, \ldots, d\}$ is a set of nodes, and E is a set of edges defined by two-element subsets of V. Two nodes u, v are *adjacent* or neighbours if uv is an edge of G and in this case the edge is drawn as bi-directed, $u \longleftrightarrow v$. Two edges are adjacent if they have an end node in common. A *path* from a node u to a node v is a sequence of adjacent edges connecting u and v for which the corresponding sequence of nodes contains no repetitions. The usual notion of *separation* in undirected graphs can be used also for bi-directed graphs; for instance, see Lauritzen (1996).

A graph G is *complete* if all its nodes are pairwise adjacent. A nonempty graph G is *connected* if any two of its nodes are linked by a path in G, otherwise it is said to be *disconnected*. If A is a subset of the node set V of G, the *induced subgraph* is denoted by G_A. If a subgraph G_A is connected (resp. disconnected, complete) we call also A connected (resp. disconnected, complete), in G. The set of all disconnected sets of the graph G will be denoted by $\mathcal{D}$, and the set of all the connected sets of G will be denoted by $\mathcal{C}$. In a graph G a *connected component* is a maximal connected subgraph. If a subset D of nodes is disconnected then it can be uniquely partitioned into its connected components $C_1, \ldots, C_r$, say, such that $D = C_1 \cup \cdots \cup C_r$.

The cardinality of a set V will be denoted by $|V|$. The set of all the subsets of V, the power set, will be denoted by $\mathcal{P}(V)$. We use also the notation $\mathcal{P}_0(V)$ for the set of all nonempty subsets of V.

Let $\boldsymbol{X} = (X_v, v \in V)$ be a discrete random vector with each component X_v taking on values in the finite set $\mathcal{I}_v = \{1, \ldots, b_v\}$. The Cartesian product $\mathcal{I}_V = \times_{v \in V} \mathcal{I}_v$, is a contingency table, with generic element $\boldsymbol{i} = (i_v, v \in V)$, called a cell of the table, and with total number of cells $t = |\mathcal{I}_V|$. We assume that $\boldsymbol{X}$ has a joint probability function $p(\boldsymbol{i})$, $\boldsymbol{i} \in \mathcal{I}_V$ giving the probability that an individual falls in cell $\boldsymbol{i}$. Given a subset $M \subseteq V$ of the variables, the marginal contingency table is $\mathcal{I}_M = \times_{v \in M} \mathcal{I}_v$ with generic cell $\boldsymbol{i}_M$ and the marginal probability function of the random vector $\boldsymbol{X}_M = (X_v, v \in M)$ is $p_M(\boldsymbol{i}_M) = \sum_{\boldsymbol{j} \in \mathcal{I}_V | \boldsymbol{j}_M = \boldsymbol{i}_M} p(\boldsymbol{j})$.

A bi-directed graph $G = (V, E)$ induces an independence model for the discrete random vector $\boldsymbol{X} = (X_v, v \in V)$ by defining a Markov property, i.e., a rule for reading off the graph the independence relations. In the following we shall use the shorthand notation $A \perp\!\!\!\perp B | C$ to indicate the conditional independence $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B | \boldsymbol{X}_C$, where A, B and C are three disjoint subsets of V. Similarly $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \perp\!\!\!\perp C$ will denote the marginal and the complete independence, respectively, of sub-vectors of $\boldsymbol{X}$. There are two Markov properties describing the independence model associated with a bi-directed graph, which we consider in this paper: (a) the global Markov property of Kauermann (1996) and (b)

the connected set Markov property by Richardson (2003).

The distribution of the random vector $\boldsymbol{X}$ satisfies the *global Markov property* for the bi-directed graph G if for any triple of disjoint sets A, B and C,

$$A \perp\!\!\!\perp B \mid V \setminus (A \cup B \cup C) \text{ whenever } A \text{ is separated from } B \text{ by } C \text{ in } G.$$

Instead, the distribution of $\boldsymbol{X}$ is said to satisfy the *connected set Markov property* if

$$\text{for every disconnected set } D \in \mathcal{D} \text{ with connected components } C_1, \ldots, C_r, \quad C_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp C_r. \quad (1)$$

Richardson (2003) proves that the two properties are equivalent; see also Drton & Richardson (2008). Following these authors we define a discrete bi-directed graph model as follows.

Definition 1. *A* discrete bi-directed graph model *associated with a bi-directed graph $G = (V, E)$ is a family of discrete joint probability distributions p for the discrete random vector $\boldsymbol{X} = (X_v, v \in V)$, that satisfies the property* (1) *for G, i.e., such that, for every disconnected set D in the graph,*

$$p_D(\boldsymbol{i}_D) = p_{C_1}(\boldsymbol{i}_{C_1}) \times \cdots \times p_{C_r}(\boldsymbol{i}_{C_r}),$$

where $C_1, \ldots, C_r$ are the connected components of D and r depends on D.

The complete list of all marginal independencies implied by a bi-directed graph model is derived from the class $\mathcal{D}$ of all disconnected sets of the graph. However, after avoiding redundancies, the same set of independencies may also be derived by considering a subclass $\mathcal{D}^* \subseteq \mathcal{D}$ containing only the *maximal disconnected* sets whose definition naturally follows from the definition of maximal connected set introduced by Richardson (2003). For the bi-directed graph in figure 1(a), the full set of independencies, i.e., $12 \perp\!\!\!\perp 4$ and $1 \perp\!\!\!\perp 34$, can be derived by considering the class $\mathcal{D}^* = \{124, 134\}$ of maximal disconnected sets or the class $\mathcal{D} = \{14, 24, 13, 124, 134\}$ of all disconnected sets. This happens because, under the connected set Markov property, the marginal independencies deriving from each non-maximal disconnected set, are necessarily implied by the independence statements induced by maximal disconnected sets.

If the global Markov property holds then for any pair of not adjacent nodes, the associated random variables are marginally independent. This implication is called the *pairwise Markov property* and it is for discrete variables a necessary but not sufficient condition for the global Markov property. This is in sharp contrast with the family of Gaussian distributions where the two properties are equivalent. In fact, the stronger condition of definition 1 implies that in some situations not all marginal independence relations are representable by bi-directed graphs. Models that define only pairwise independencies but not joint independencies appear to be of very limited interest. However, the following example is an exception in this respect.

Example 1. Consider the data in table 1, due to Lienert (1970) and discussed by Wermuth (1998). The variables are 3 symptoms after LSD intake, recorded to be present (level 1) or absent (level 2), and are distortions in affective behavior (X_1), distortions in thinking (X_2), and dimming of consciousness (X_3). The frequencies in the three marginal tables

show that the three symptom pairs are close to independence, but at the same time the variables are not mutually independent as witnessed by the strong three-factor interaction due to the quite distinct conditional odds ratios between X_1 and X_2 at the two levels of X_3. Thus, in this case, despite three marginal independencies, a discrete bi-directed graph model can represent just one of them, and thus must include at least two edges.

[Table 1 about here.]

3 Marginal log-linear parameterizations

Discrete bi-directed graph models may be defined as marginal log-linear models, using complete hierarchical parameterizations by Bergsma & Rudas (2002). In this section we review the basic concepts and we discuss the definitions of the parameters involved. Let $p(\boldsymbol{i}) > 0$ be a strictly positive probability distribution of a discrete random vector $\boldsymbol{X} = (X_v, v \in V)$ and let $p_M(\boldsymbol{i}_M)$ be any marginal probability distribution of a sub-vector $\boldsymbol{X}_M$, $M \subseteq V$. The marginal probability distribution admits a log-linear expansion

$$\log p_M(\boldsymbol{i}_M) = \sum_{L \subseteq M} \lambda_L^M(\boldsymbol{i}_L)$$

where $\lambda_L^M(\boldsymbol{i}_L)$ is a function defining the log-linear parameters indexed by the subset L of M. The functions $\lambda_L^M(\boldsymbol{i}_L)$ are defined by

$$\lambda_L^M(\boldsymbol{i}_L) = \sum_{A \subseteq L} (-1)^{|L \setminus A|} \log p_M(\boldsymbol{i}_A, \boldsymbol{i}^*_{M \setminus A})$$

where $\boldsymbol{i}^* = (1, \ldots, 1)$ denotes a baseline cell of the table; see Whittaker (1990, p. 206) and Lauritzen (1996, App. B.2, p. 249). The function $\lambda_L^M(\boldsymbol{i}_L)$ is zero whenever at least one index in $\boldsymbol{i}_L$ is equal to 1. Therefore, $\lambda_L^M(\boldsymbol{i}_L)$ defines only $\prod_{v \in L}(b_v - 1)$ parameters where b_v is the number of categories of variable X_v. Due to the constraint on the probabilities, that must sum to one, the parameter $\lambda_\phi^M = \log p(\boldsymbol{i}^*_M)$ is a function of the others, and can thus be eliminated.

If $\boldsymbol{\lambda}_L^M$ is the vector containing the parameters $\boldsymbol{\lambda}_L^M(\boldsymbol{i}_L)$, then it can be obtained explicitly using Kronecker products as follows. For any subset L of M, let $\boldsymbol{C}_{v,L}$ be the matrix

$$\boldsymbol{C}_{v,L} = \begin{cases} (-\mathbf{1}_{b_v - 1} \quad \boldsymbol{I}_{b_v - 1}) & \text{if } v \in L \\ (1 \quad \mathbf{0}_{b_v - 1}) & \text{if } v \notin L. \end{cases}$$

and let $\boldsymbol{\pi}^M$ be the $t_M \times 1$ column vector of the marginal cell probabilities in lexicographic order. Then, the vector of the log-linear parameters $\boldsymbol{\lambda}_L^M(\boldsymbol{i}_L)$ is

$$\boldsymbol{\lambda}_L^M = \boldsymbol{C}_L^M \log \boldsymbol{\pi}^M, \text{ where } \boldsymbol{C}_L^M = \bigotimes_{v \in M} \boldsymbol{C}_{v,L}. \tag{2}$$

The above definition corresponds to the baseline contrasts, that are the default choice of the R environment (R Development Core Team, 2008); see also Wermuth & Cox (1992).

A marginal log-linear parameterization of the probability distribution $p(\boldsymbol{i})$ is obtained by combining the log-linear parameters $\boldsymbol{\lambda}_L^M$ for many different marginal probability distributions. The general theory is developed in Bergsma & Rudas (2002) and is summarized below.

Definition 2. *Let $\mathcal{M} = (M_1, \ldots, M_s)$ be an ordered sequence of margins of interest, and, for each M_j, $j = 1, \ldots, s$, let $\mathcal{L}_j$ be the collection of sets L for which $\boldsymbol{\lambda}_L^{M_j}$ is defined by equation (2). Then, $(\boldsymbol{\lambda}_L^{M_j} | L \in \mathcal{L}_j, j = 1, \ldots, s)$ is said to be a hierarchical and complete marginal log-linear parameterization for $p(\boldsymbol{i})$ if (i) the sequence $M_1, \ldots, M_s$ is nondecreasing: $M_i \nsubseteq M_j$ if $i > j$; (ii) the last margin is $M_s = V$; (iii) the sets defining the log-linear parameters in each margin are:*

$$\mathcal{L}_1 = \mathcal{P}_0(M_1), \text{ and } \mathcal{L}_j = \mathcal{P}_0(M_j) \setminus \bigcup_{h=1}^{j-1} \mathcal{L}_h, \text{ for } j > 1,$$

where $\mathcal{P}_0(M_j)$ denotes the collection of all nonempty subsets of M_j.

The parameterization is called hierarchical because the log-linear interactions defined in each margin form an ascending class and complete because it defines all possible log-linear terms, each within one and only one marginal table. Notice that each nondecreasing sequence $\mathcal{M}$ of margins defines a unique parameterization. Thus also a reordering of the sequence that preserves its properties yields a different parameterization; see the examples in section 4.3.

The above construction defines a map from the simplex Δ_V of the strictly positive distributions $p(\boldsymbol{i})$ of the discrete random vector $\boldsymbol{X}$ into the set Λ of possible values for the whole vector of the marginal log-linear parameters $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_L^{M_j})$, with $j = 1, \ldots, s$ and $L \in \mathcal{L}_j$. The following general result shows that a complete hierarchical marginal log-linear model defines a proper parameterization.

Proposition 1. *(Bergsma & Rudas, 2002) The map $\Delta_V \to \Lambda \subseteq \mathbf{R}^{t-1}$ defined by a complete and hierarchical marginal log-linear parameterization is a diffeomorphism.*

The parameters $\boldsymbol{\lambda}$ can be written in matrix form

$$\boldsymbol{\lambda} = \boldsymbol{C} \log(\boldsymbol{T}\boldsymbol{\pi})$$

where $\boldsymbol{\pi}$ is the $t \times 1$ vector of all the cell probabilities in lexicographical order, $\boldsymbol{T}$ is a $m \times t$ marginalization matrix such that

$$\boldsymbol{T}\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}^{M_1} \\ \vdots \\ \boldsymbol{\pi}^{M_s} \end{pmatrix}$$

and $\boldsymbol{C} = \text{diag}(\boldsymbol{C}_L^M)$ is a $t - 1 \times m$ block diagonal matrix, with $m = \sum_{j=1}^{s} |\mathcal{I}_{M_j}|$. For a discussion of algorithms for computing the matrices $\boldsymbol{C}$ and $\boldsymbol{T}$ see Bartolucci *et al.* (2007), who generalize the approach by Bergsma & Rudas (2002) to logits and higher order effects of global and continuation type, suitable for ordinal data .

The log-linear parameterization and the multivariate logistic transformation are two special cases of marginal log-linear models. The overall log-linear parameters are generated by $\mathcal{M} = \{V\}$. They will be denoted by $\boldsymbol{\theta}_L = \boldsymbol{\lambda}_L^V$ for $L \in \mathcal{P}_0(V)$ and the whole vector of parameters by $\boldsymbol{\theta}$. The parameter space coincides with $\mathbf{R}^{t-1}$ and the map from $\boldsymbol{\pi}$ to $\boldsymbol{\theta}$ admits an inverse in closed form, provided that $\boldsymbol{\pi} > \mathbf{0}$. The multivariate logistic

parameters of Glonek & McCullagh (1995) are generated by $\mathcal{M} = \mathcal{P}_0(V)$, in any nondecreasing order. They will be denoted by $\boldsymbol{\eta}^M = \boldsymbol{\lambda}^M_M$, with $\boldsymbol{\eta}$ representing the whole vector. Thus the parameters $\boldsymbol{\eta}^M$ correspond to the highest order log-linear parameters within each marginal table $\mathcal{I}_M$, for each nonempty set $M \subseteq V$. The parameter space is in general a strict subset of $\mathbf{R}^{t-1}$, except when the number of variables is $d = 2$. In general there is no closed form inverse transforming back $\boldsymbol{\eta}$ into $\boldsymbol{\pi}$. The inverse operation however may be accomplished using for example the iterative proportional fitting algorithm or other methods; for a recent discussion see Qaqish & Ivanova (2006).

Thus, while the log-linear parameters $\boldsymbol{\theta}$ are always variation independent and for any $\boldsymbol{\theta}$ in $\mathbf{R}^{t-1}$ there is a unique associated joint probability distribution $\boldsymbol{\pi}$, instead the multivariate logistic parameters are never variation independent, for $d > 2$. Thus there are vectors $\boldsymbol{\eta}$ in $\mathbf{R}^{t-1}$ that are not compatible with any joint probability distribution $\boldsymbol{\pi}$. The latter assertion is also implied by a further result by Bergsma & Rudas (2002) which proves that the hierarchical and complete marginal log-linear parameterization generated by a sequence $\mathcal{M}$ is variation independent if and only if $\mathcal{M}$ satisfies a property called *ordered decomposability*. A sequence of arbitrary subsets of V is said to be ordered decomposable if it has at most two elements or if there is an ordering $M_1, \ldots, M_s$ of its elements, such that $M_i \not\subseteq M_j$ if $i > j$ and, for $k = 3, \ldots, s$, the maximal elements (i.e., those not contained in any other sets) of $\{M_1, \ldots, M_k\}$ form a decomposable set. For further details and examples about ordered decomposability see Rudas & Bergsma (2004). More properties of the two parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, connected to graphical models, will be described in the next section 4.

4 Parameterizations of discrete bi-directed graph models

We suggest now two different marginal log-linear parameterizations of discrete bi-directed graph models, and we compare advantages and shortcomings.

4.1 Multivariate logistic parameterization

It is known that the complete independence of two sub-vectors $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ of the random vector $\boldsymbol{X}$ is equivalent to a set of zero restrictions on multivariate logistic parameters.

Lemma 1. (Kauermann (1997), lemma 1). *If $\{A, B\}$ is a partition of V and $\boldsymbol{\eta} = (\boldsymbol{\eta}^M), M \in \mathcal{P}_0(V)$ is the multivariate logistic parameterization, then*

$$A \perp\!\!\!\perp B \iff \boldsymbol{\eta}^M = \mathbf{0} \quad \text{for all } M \in \mathcal{Q}$$

where $\mathcal{Q} = \{M \subseteq A \cup B : M \cap A \neq \emptyset, M \cap B \neq \emptyset\}$.

We generalize this result to complete independence of more than two random vectors. Given a partition $\{C_1, \ldots, C_r\}$ of a set $D \subseteq V$, we define

$$\mathcal{Q}(C_1, \ldots, C_r) = \mathcal{P}\left(\bigcup_{i=1}^r C_k\right) \setminus \bigcup_{i=1}^r \mathcal{P}(C_k).$$

This is the set of all subsets of D not completely contained in a single class, i.e., containing elements coming from at least two classes of the partition. With this notation, the set $\mathcal{Q}$ of lemma 1 may be denoted by $\mathcal{Q}(A, B)$. Then we have the following result.

Proposition 2. *Let $\boldsymbol{X} = (X_v), v \in V$, be the discrete random vector with multivariate logistic parameterization $\boldsymbol{\eta} = (\boldsymbol{\eta}^M), M \in \mathcal{P}_0(V)$. If $D \subseteq V$ is partitioned into the classes $\{C_1, \ldots, C_r\}$ then*

$$C_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp C_r \iff \text{for all } M \in \mathcal{Q}(C_1, \ldots, C_r): \quad \boldsymbol{\eta}^M = \mathbf{0}.$$

Proof. First, use the shorthand notations $\mathcal{Q}$ to denote the set $\mathcal{Q}(C_1, \ldots, C_r)$ and $\mathcal{Q}_i$ to denote the set $\mathcal{Q}(C_i, C_{-i})$, $i = 1, \ldots, r$, where $C_{-i} = D \setminus C_i$. In fact, since $\mathcal{Q}_i \subseteq \mathcal{Q}$, then $\bigcup_{i=1}^r \mathcal{Q}_i \subseteq \mathcal{Q}$. Conversely, for any $M \in \mathcal{Q}$ there is always a class C_i such that $C_i \nsubseteq M$, and hence, by definition, $M \in \mathcal{Q}_i$. Hence, for every $M \in \mathcal{Q}$, $M \in \bigcup_{i=1}^r \mathcal{Q}_i$ and thus $\mathcal{Q} \subseteq \bigcup_{i=1}^r \mathcal{Q}_i$. Then, the complete independence $C_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp C_r$ is equivalent to $C_i \perp\!\!\!\perp C_{-i}$ for all $i = 1, \ldots, r$. By lemma 1, applied to the sub-vector $\boldsymbol{X}_D$, each independence $C_i \perp\!\!\!\perp C_{-i}$ is equivalent to the restriction $\boldsymbol{\eta}^M = \mathbf{0}$ for $M \in \mathcal{Q}_i$ and the parameters $\boldsymbol{\eta}^M$ are identical to the corresponding multivariate logistic parameters for the full random vector $\boldsymbol{X}_V$. Thus, the complete independence $C_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp C_r$ is equivalent to $\boldsymbol{\eta}^M = \mathbf{0}$ for $M \in \mathcal{Q}_i$, $i = 1, \ldots, r$, i.e., for $M \in \bigcup_{i=1}^r \mathcal{Q}_i = \mathcal{Q}$.

Proposition 2 implies that a statement of complete independence $C_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp C_r$ is equivalent to a set of zero constraints on the multivariate logistic parameters. The following result explains how the constraints must be chosen in order to satisfy all the independencies required by definition 1 of a bi-directed graph model.

Proposition 3. *Given a bi-directed graph $G = (V, E)$, the discrete bi-directed graph model associated with G is defined by the set of strictly positive discrete probability distributions with multivariate logistic parameters $\boldsymbol{\eta} = (\boldsymbol{\eta}^M)$, $M \in \mathcal{P}_0(V)$, such that*

$$\boldsymbol{\eta}^M = \mathbf{0} \text{ for every } M \in \mathcal{D},$$

where $\mathcal{D}$ is the set of all disconnected sets of nodes in the graph G.

Proof. Given a set $D \in \mathcal{D}$, denote its connected components by $\{C_1, \ldots C_r\}$ and by $\mathcal{Q}_D$ the set $\mathcal{Q}(C_1, \ldots, C_r)$. First, we prove that $\mathcal{D} = \bigcup_{D \in \mathcal{D}} \mathcal{Q}_D$. In fact, for any $D \in \mathcal{D}$, $\mathcal{Q}_D \subseteq \mathcal{D}$ because it is a class of disconnected subsets of D. Thus, $\bigcup_{D \in \mathcal{D}} \mathcal{Q}_D \subseteq \mathcal{D}$. Conversely, if $D \in \mathcal{D}$, then $D \in \mathcal{Q}_D$ and thus $\mathcal{D} \subseteq \bigcup_{D \in \mathcal{D}} \mathcal{Q}_D$. By definition 1, the independence $C_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp C_r$ is implied for each disconnected set D with connected components $C_1, \ldots, C_r$. By proposition 2, this is equivalent to the zero restrictions on the multivariate logistic parameters

$$\boldsymbol{\eta}^M = \mathbf{0}, \text{ for all } M \in \mathcal{Q}_D, \quad D \in \mathcal{D}$$

i.e., for all $M \in \bigcup_{D \in \mathcal{D}} \mathcal{Q}_D = \mathcal{D}$.

[Table 2 about here.]

A consequence of proposition 3 is that all possible discrete bi-directed graphical models can be identified within the multivariate logistic parameterization under the zero constraints associated with the disconnected sets.

Example 2. The discrete model associated with the chordless 4-chain of figure 1(a) is defined by the multivariate logistic parameters shown in table 2, first row. There are 5

zero constraints on the highest-order log-linear parameters of the tables 13, 14, 24, 124 134. There are three nonzero two-factor marginal log-linear parameter $\boldsymbol{\eta}^{ij}$ associated with the edges of the graph that may be interpreted as sets of marginal association coefficients between the involved variables, based on the chosen contrasts. Consider now the reduced model resulting after dropping the edge $2 \leftrightarrow 3$ and implying the independence $12 \perp\!\!\!\perp 34$. This model can be obtained, within the same parameterization, by the additional zero constraints on $\boldsymbol{\eta}^{23}, \boldsymbol{\eta}^{123}, \boldsymbol{\eta}^{234}$ and $\boldsymbol{\eta}^{1234}$.

While the parameters are in general not variation independent, they satisfy the upward compatibility property, because they have the same meaning across different marginal distributions. Using this property, we can prove the following result concerning the effect of marginalization over a subset A of the variables. Let $G_A = (A, E_A)$ be the subgraph induced by A, and let $\mathcal{D}_A$ be the set of all disconnected sets of G_A.

Proposition 4. *If a discrete probability distribution $p(\boldsymbol{i})$ for $\boldsymbol{i} \in \mathcal{I}_V$ satisfies a bi-directed graph model defined by the graph $G = (V, E)$ then the marginal distribution $p_A(\boldsymbol{i}_A)$ over $A \subseteq V$ satisfies the bi-directed graph model defined by $G_A = (A, E_A)$ and its multivariate logistic parameters are $\boldsymbol{\eta} = (\boldsymbol{\eta}^M), M \in \mathcal{P}_0(A)$ with constraints $\boldsymbol{\eta}^M = \mathbf{0}$, for $M \in \mathcal{D}_A$.*

Proof. After marginalization over A, the multivariate logistic parameters associated with $p_A(\boldsymbol{i}_A)$, by the property of upward compatibility, are $(\boldsymbol{\eta}^M, M \in \mathcal{P}_0(A))$. Some of these parameters are zero by the constraints implied by the original bi-directed graph model, i.e., $\boldsymbol{\eta}^M = \mathbf{0}$, for $M \in \mathcal{D} \cap \mathcal{P}_0(A)$. The result is proved by showing that $\mathcal{D} \cap \mathcal{P}_0(A) = \mathcal{D}_A$. First, we note that if $D \subseteq A \subseteq V$, then the graph $G_D = (D, E_D)$ with edges $E_D = (D \times D) \cap E = (D \times D) \cap E_A$ is a subgraph of both G_A and G. Thus, if $D \subseteq A$ and $D \in \mathcal{D}$ then the induced subgraph G_D is disconnected and being also a subgraph of G_A then D is also a disconnected set of G_A. Thus $\mathcal{D} \cap \mathcal{P}_0(A) \subseteq \mathcal{D}_A$. Conversely, if D is a disconnected set of G_A, then the subgraph G_D is disconnected, and being a subgraph of G, then D is also a disconnected set of G. Thus $\mathcal{D}_A \subseteq \mathcal{D} \cap \mathcal{P}(A)$, and the result follows.

Discrete bi-directed graph models in the multivariate logistic parameterization can be compared with discrete log-linear graphical models represented by undirected graphs with the same skeleton (i.e., with the same set E). In the second row of table 2, we show the log-linear parameterization $\boldsymbol{\theta}$ for a discrete undirected 4-chain graph model, with independencies $12 \perp\!\!\!\perp 4|3$ and $1 \perp\!\!\!\perp 34|2$. From the Hammersley-Clifford theorem, (see Lauritzen, 1996, p. 36), this model is defined by setting to zero the log-linear interactions indexed by the eight incomplete subsets of the graph. As stated by Drton & Richardson (2008), the number of zero restrictions of an undirected graph model is always higher compared to that of the bi-directed graph model with the same skeleton, because the set of incomplete subsets necessarily includes all disconnected sets.

In discrete undirected graph models the general hierarchy principle holds; for instance, from table 2, $\boldsymbol{\theta}_{13} = \mathbf{0}$, and also $\boldsymbol{\theta}_{123} = \boldsymbol{\theta}_{134} = \boldsymbol{\theta}_{1234} = \mathbf{0}$. In contrast, in the multivariate logistic parameterization of the bi-directed graph model the hierarchy principle is violated. Thus, as shown in table 2 there are zero pairwise associations, like $\boldsymbol{\eta}^{13} = \mathbf{0}$, but nonzero higher order log-linear parameters, e.g., $\boldsymbol{\eta}^{123} \neq \mathbf{0}$ and $\boldsymbol{\eta}^{1234} \neq \mathbf{0}$.

4.2 The disconnected sets parameterization

We discuss now another marginal log-linear parameterization that can represent the independence constraints implied by any discrete bi-directed graph model, but involving only those marginal tables which are needed. This parameterization defines the log-linear parameters within the margins associated with the disconnected sets of the graph defining the model. Specifically, given a discrete graph model with a graph G, we arbitrarily order the disconnected sets of the graph to yield a nondecreasing sequence $(D_1, \ldots, D_s)$ such that $D_k \not\supseteq D_{k+1}$ for $k = 1, \ldots, s-1$. Then, the *disconnected set parameterization* of the discrete bi-directed graph model associated with G, is the hierarchical and complete marginal log-linear parameterization $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_L^{M_j})$ generated, following definition 2, by the sequence of margins

$$\mathcal{M}_G = \begin{cases} (D_1, \ldots, D_s) & \text{if } D_s = V \\ (D_1, \ldots, D_s, V) & \text{otherwise.} \end{cases} \tag{3}$$

This parameterization contains by definition the log-linear parameters $\boldsymbol{\lambda}_D^D = \boldsymbol{\eta}^D$ for every disconnected set D and thus can define the independence model by the same constraints of proposition 3.

Proposition 5. *Given a bi-directed graph $G = (V, E)$, the discrete bi-directed graph model associated with G is defined by the set of strictly positive discrete probability distributions with a disconnected set parameterization $(\boldsymbol{\lambda}_L^{M_j})$, such that*

$$\boldsymbol{\lambda}_{M_j}^{M_j} = \mathbf{0} \text{ for every } M_j \in \mathcal{D},$$

where $\mathcal{D}$ is the class of all disconnected sets for G. Moreover, the constraints are independent of the ordering chosen to define $\mathcal{M}_G$.

Proof. The disconnected set parameterization defined by the sequence (3), contains the parameters $\boldsymbol{\lambda}_L^D$, with $D \in \mathcal{D}$. By definition 2, $\mathcal{L}_j$, $j = 1, \ldots, s$ always contains the set D itself. This happens whatever ordering is used to define $\mathcal{M}_G$. Thus the parameterization always includes $\boldsymbol{\lambda}_D^D = \boldsymbol{\eta}^D$, for every $D \in \mathcal{D}$ and it is possible to impose the constraints $\boldsymbol{\eta}^D = \mathbf{0}$ for every $D \in \mathcal{D}$ and the result follows by proposition 3.

While the constrained parameters defining the bi-directed graph model are actually the same as the multivariate logistic parameterization, the other unconstrained log-linear parameters are defined in larger marginal tables, and thus have a different interpretation. An important difference is that the disconnected set parameterization is tied to the specific graph G defining the model. This implies that it is not possible to define every bi-directed graph model within the same disconnected set parameterization. A different model G implies a different sequence $\mathcal{M}_G$ of disconnected sets and thus a different list of log-linear parameters.

Example 3. For the chordless 4-chain graph of figure 1(a), there are several possible orderings of the 5 disconnected sets $\mathcal{D} = \{13, 14, 24, 134, 124\}$. The discrete bi-directed graph model is defined by choosing for example

$$\mathcal{M}_G = (13, 14, 24, 134, 124, 1234),$$

and by constraining the marginal log-linear parameters $\boldsymbol{\lambda}_D^D = \mathbf{0}$ for $D \in \mathcal{D}$. The unconstrained parameters differ from the multivariate logistic ones. For example the two-factor log-linear parameters for X_1 and X_2, $\boldsymbol{\lambda}_{12}^{124}$, are defined within the margin 124 instead of the margin 12. A detailed comparison of the parameters is reported in the first two rows of table 3. Notice that the parameters are defined by using a reduced number of marginal tables.

[Table 3 about here.]

4.3 Comparison of different parameterizations

We now show that a discrete bi-directed graph model can also be defined by zero restrictions on parameters different from the multivariate logistic ones, but still indexed by the disconnected sets. We limit the discussion to the case of 4-chain. We propose two alternative selections of marginal tables: the first suggested by the global Markov property, the second by the connected set Markov property for the maximal disconnected sets. Although we have no formal proof, computational results suggest that these alternative parameterizations fulfill the independence model for every discrete bi-directed graph model.

Example 4. For the 4-chain graph of figure 1(a), the global Markov property implies the conditional independencies $1 \perp\!\!\!\perp 4$, $2 \perp\!\!\!\perp 4|1$ and $1 \perp\!\!\!\perp 3|4$. Thus, the relevant margins can be collected in the sequence

$$\mathcal{M}'_G = (14, 134, 124, 1234)$$

where the first three allow the definition of the conditional independencies and the last one serves as completion of the parameterization. The complete hierarchical parameterization generated by $\mathcal{M}'_G$ is slightly different from that generated by $\mathcal{M}_G$, see table 3, third row, but with the 5 zero constraints on the higher level log-linear parameters within each margin, we obtain the required independencies

$$1 \perp\!\!\!\perp 4 \iff \boldsymbol{\lambda}_{14}^{14} = \mathbf{0} \quad 1 \perp\!\!\!\perp 3|4 \iff \begin{cases} \boldsymbol{\lambda}_{13}^{134} = \mathbf{0} \\ \boldsymbol{\lambda}_{134}^{134} = \mathbf{0} \end{cases} \qquad 2 \perp\!\!\!\perp 4|1 \iff \begin{cases} \boldsymbol{\lambda}_{24}^{124} = \mathbf{0} \\ \boldsymbol{\lambda}_{124}^{124} = \mathbf{0}. \end{cases}$$

Note that these independencies can also be represented by a chain graph with two components, $\{1, 4\}$ and $\{2, 3\}$, under the multivariate regression Markov property; see Marchetti & Lupparelli (2008) and Drton (2008). The associated discrete model is interpreted as a system of seemingly unrelated regressions, with two joint responses X_2 and X_3 and two independent explanatory variables X_1 and X_4.

Example 5. For the 4-chain graph of figure 1(a), the connected set Markov property implies the marginal independencies $1 \perp\!\!\!\perp 34$ and $12 \perp\!\!\!\perp 4$. Thus, we consider the sequence

$$\mathcal{M}''_G = (134, 124, 1234)$$

including only the maximal disconnected sets 134 and 124. The complete hierarchical parameterization generated by $\mathcal{M}''_G$ is shown in table 3, fourth row. We obtain the required

independencies because the independencies $1 \perp\!\!\!\perp 34$ and $12 \perp\!\!\!\perp 4$ are equivalent to $1 \perp\!\!\!\perp 34$ and $2 \perp\!\!\!\perp 4 \mid 1$, and

$$1 \perp\!\!\!\perp 34 \iff \begin{cases} \boldsymbol{\lambda}_{13}^{134} = \mathbf{0} \\ \boldsymbol{\lambda}_{14}^{134} = \mathbf{0} \\ \boldsymbol{\lambda}_{134}^{134} = \mathbf{0} \end{cases} \qquad 2 \perp\!\!\!\perp 4 \mid 1 \iff \begin{cases} \boldsymbol{\lambda}_{24}^{124} = \mathbf{0} \\ \boldsymbol{\lambda}_{124}^{124} = \mathbf{0}. \end{cases}$$

The required independencies $1 \perp\!\!\!\perp 34$ and $12 \perp\!\!\!\perp 4$ are also equivalent to $12 \perp\!\!\!\perp 4$ and $1 \perp\!\!\!\perp 3 \mid 4$ which can be obtained by constraining the same interaction terms in the parameterization generated by the reordered sequence $\mathcal{M}_G = (124, 134, 1234)$.

[Figure 2 about here.]

Given a bi-directed graph, there is a whole class of hierarchical and complete marginal log-linear parameterizations that involve parameters for disconnected sets and have the property that setting the parameters for disconnected sets to zero yields the distributions in the considered bi-directed graph model. As indicated by a referee, this class of models could be ordered according to the partial order defined in Bergsma & Rudas (2002, p. 143), with the multivariate logistic parameterization representing the unique minimal element. For instance, the different parameterizations discussed for a 4-chain and illustrated in table 3 are naturally ordered as $\eta \ll \mathcal{M}_G \ll \mathcal{M}'_G \ll \mathcal{M}''_G$. Finding the maximal element of this class is an open problem. We conjecture that a maximal element is represented by the parameterizations generated by the nondecreasing sequences of maximal disconnected sets.

In the comparison of different parameterizations the property of variation independence may also relevant. Following Bergsma & Rudas (2002), there is a variation independent parameterization if there is at least one sequence $\mathcal{M}_G$ which is ordered decomposable. This property is quite relevant because the lack of variation independence may make the separate interpretation of the parameters misleading.

Example 6. In examples 3, 4 and 5, the parameterizations based on $\mathcal{M}_G$, $\mathcal{M}'_G$ and $\mathcal{M}''_G$ are variation independent (unlike the multivariate logistic parameterization) because the sequences of margins are all ordered decomposable. Consider instead the bi-directed graph in figure 2(a). Two possible disconnected set parameterizations of the discrete model may be based for example on

$$\begin{aligned} \mathcal{M}_G &= (13, 14, 25, 35, 134, 135, 235, 12345), \\ \mathcal{M}'_G &= (13, 35, 135, 14, 25, 134, 235, 12345). \end{aligned}$$

with the constraints $\boldsymbol{\lambda}_D^D = \mathbf{0}$ for any disconnected set D. In this case we can verify that only the sequence $\mathcal{M}'_G$ is ordered decomposable and thus implies variation independent parameters.

The identification of the class of bi-directed graph models admitting a variation independent parameterization is an open problem. Below, we provide some partial results with respect to this issue.

Variation independence is always achieved for bi-directed graphs whose disconnected sets are disjoint. For instance, let us consider the chordless 4-cycle in figure 1(b). The disconnected sets 13 and 24 are disjoint and the parameterization is variation independent whichever the chosen ordering of $\mathcal{M}_G$. In this case, the variation independence is also trivially achieved by definition because there are only two disconnected sets.

An exhaustive search among all bi-directed graphs with three and four nodes shows that the disconnected set parameterization is always variation independent whichever the chosen ordering $\mathcal{M}_G$.

With five nodes, example 6 shows that this property depends on the chosen ordering of margins. The first difficult situation arises for the chordless 5-cycle and the 5-chain. In these cases the disconnected set parameterizations are never variation independent, neither in the basic version nor in their maximal version discussed in section 4.3. Following Drton & Richardson (2008), we conjecture that complete hierarchical variation independent parameterizations do not exist for these two cases. All the previous examples indicate that algorithms for finding an order decomposable sequence of disconnected sets are important (whenever it exists) and need to be investigated.

5 Maximum likelihood estimation of discrete bi-directed graph models

We study now the maximum likelihood estimation of the discrete bi-directed graph models under any of the parameterizations previously discussed. Assuming a multinomial sampling scheme with sample size N, each individual falls in a cell $\boldsymbol{i}$ of the given contingency table $\mathcal{I}_V$ with probability $p(\boldsymbol{i}) > 0$. Let $n(\boldsymbol{i})$ be the cell count and $\boldsymbol{n} = (n(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}_V)$, be a $t \times 1$ vector. Thus, $\boldsymbol{n}$ has a multinomial distribution with parameters N and $\boldsymbol{\pi}$. If $\boldsymbol{\mu} = N\boldsymbol{\pi} > \mathbf{0}$ is the expected value of $\boldsymbol{n}$ and $\boldsymbol{\omega} = \log \boldsymbol{\mu}$, then for any appropriate marginal log-linear parameterization $\boldsymbol{\lambda}$ we have $\boldsymbol{\lambda} = \boldsymbol{C} \log(\boldsymbol{T}\boldsymbol{\pi}) = \boldsymbol{C} \log(\boldsymbol{T} \exp(\boldsymbol{\omega}))$ because the contrasts of marginal probabilities are equal to the contrasts of expected counts. Given a discrete bi-directed graph model defined by the graph $G = (V, E)$, if $\boldsymbol{\lambda}$ is defined either by the multivariate logistic parameterization or by the disconnected set parameterization, we can always split $\boldsymbol{\lambda}$ in two components $\boldsymbol{\lambda}_{\mathcal{D}}$ and $\boldsymbol{\lambda}_{\mathcal{C}}$ indexed by the disconnected sets $\mathcal{D}$ and by the connected sets $\mathcal{C}$ of the graph, respectively. If $\boldsymbol{C}_{\mathcal{D}}$ is a sub-matrix of the contrast matrix $\boldsymbol{C}$, obtained by selecting the rows associated with the disconnected sets of the graph G,

$$\boldsymbol{\lambda}_{\mathcal{D}} = \boldsymbol{C}_{\mathcal{D}} \log(\boldsymbol{T} \exp(\boldsymbol{\omega})) = \boldsymbol{h}(\boldsymbol{\omega})$$

where $\boldsymbol{C}_{\mathcal{D}}$ has dimensions $q \times v$ with $q = \sum_{D \in \mathcal{D}} \prod_{v \in D} (b_v - 1)$. Thus, the kernel of the log-likelihood function of the discrete bi-directed graph model is defined by

$$l(\boldsymbol{\omega}; \boldsymbol{n}) = \boldsymbol{n}^{\mathrm{T}} \boldsymbol{\omega} - \mathbf{1}^{\mathrm{T}} \exp(\boldsymbol{\omega}), \quad \boldsymbol{\omega} \in \Omega_{BG}, \tag{4}$$

with

$$\Omega_{BG} = \{\boldsymbol{\omega} \in \mathbf{R}^t : \boldsymbol{h}(\boldsymbol{\omega}) = \mathbf{0}, \quad \mathbf{1}^{\mathrm{T}} \exp(\boldsymbol{\omega}) = N\}.$$

Note that (4) defines a curved exponential family model as the set Ω_{BG} is a smooth manifold in the space $\mathbf{R}^t$ of the canonical parameters $\boldsymbol{\mu}$. Maximum likelihood estimation

is a constrained optimization problem and the maximum likelihood estimate is a saddle point of the Lagrangian log-likelihood

$$\ell(\boldsymbol{\omega}, \boldsymbol{\tau}) = \boldsymbol{n}^{\mathrm{T}}\boldsymbol{\omega} - \mathbf{1}^{\mathrm{T}} \exp(\boldsymbol{\omega}) + \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{\omega})$$

where $\boldsymbol{\tau}$ is a $q \times 1$ vector of unknown Lagrange multipliers. To solve the equations we propose an iterative procedure inspired by Aitchison & Silvey (1958), Lang (1996) and Bergsma (1997). Define first

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\omega} \\ \boldsymbol{\tau} \end{pmatrix}, \quad \boldsymbol{f}(\boldsymbol{\xi}) = \frac{\partial \ell}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \boldsymbol{f}_{\omega} \\ \boldsymbol{f}_{\tau} \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{\xi}) = -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^{\mathrm{T}}}\right) = \begin{pmatrix} \boldsymbol{F}_{\omega\omega} & \boldsymbol{F}_{\omega\tau} \\ \cdot & \boldsymbol{F}_{\tau\tau} \end{pmatrix},$$

where the dot is a shortcut to denote a symmetric sub-matrix. Differentiating the Lagrangian with respect to $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$ and equating the result to zero we obtain

$$\begin{pmatrix} \boldsymbol{f}_{\omega} \\ \boldsymbol{f}_{\tau} \end{pmatrix} = \begin{pmatrix} \boldsymbol{e} + \boldsymbol{H}\boldsymbol{\tau} \\ \boldsymbol{h}(\boldsymbol{\omega}) \end{pmatrix} = \mathbf{0} \tag{5}$$

where $\boldsymbol{e} = \partial l / \partial \boldsymbol{\omega} = \boldsymbol{n} - \boldsymbol{\mu}$, $\boldsymbol{H} = \partial \boldsymbol{h} / \partial \boldsymbol{\omega}^{\mathrm{T}} = \boldsymbol{D}_{\mu}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{D}_{T\mu}^{-1}\boldsymbol{C}_{\mathcal{D}}^{\mathrm{T}}$ and $\boldsymbol{D}_{T\mu}$ and $\boldsymbol{D}_{\mu}$ are diagonal matrices, with nonzero elements $\boldsymbol{T}\boldsymbol{\mu}$ and $\boldsymbol{\mu}$, respectively.

Let $\hat{\boldsymbol{\omega}}$ be a local maximum of the likelihood subject to the constraint $\boldsymbol{h}(\boldsymbol{\omega}) = \mathbf{0}$. A classical result (Bertsekas, 1982) is that if $\boldsymbol{H}$ is of full column rank at $\hat{\boldsymbol{\omega}}$, there is a unique $\hat{\boldsymbol{\tau}}$ such that $\ell(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\tau}}) = \mathbf{0}$. In the sequel, it is assumed that the maximum likelihood estimate $\hat{\boldsymbol{\omega}}$ is a solution to the equation (5). Note that the constraint $\mathbf{1}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{1}^{\mathrm{T}}\boldsymbol{n}$ is automatically satisfied as it can be verified that $\boldsymbol{H}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$ and thus from (5) it follows that $\mathbf{1}^{\mathrm{T}}\boldsymbol{e} = \mathbf{0}$.

Aitchison and Silvey propose a Fisher score like updating function

$$\boldsymbol{\xi}^{(k+1)} = \boldsymbol{u}(\boldsymbol{\xi}^{(k)}), \text{ with } \boldsymbol{u}(\boldsymbol{\xi}) = \boldsymbol{\xi} + \boldsymbol{F}^{-1}(\boldsymbol{\xi})\boldsymbol{f}(\boldsymbol{\xi}), \tag{6}$$

yielding the estimate $\boldsymbol{\xi}^{(k+1)}$ at cycle $k+1$ from that at cycle k. As the algorithm does not always converge when starting estimates are not close enough to $\hat{\boldsymbol{\omega}}$, it is necessary to introduce a step size into the updating equation. The standard approach to choosing a step size in optimization problems is to use a value for which the objective function to be maximized increases. However, since in in this case we are looking for a saddle point of the Lagrangian likelihood ℓ, we need to adjust the standard strategy. First, the matrix $\boldsymbol{F}$ has a special structure with $\boldsymbol{F}_{\omega\omega} = \boldsymbol{D}_{\mu}$, $\boldsymbol{F}_{\omega\tau} = -\boldsymbol{H}$ and $\boldsymbol{F}_{\tau\tau} = \mathbf{0}$. Thus, indicating the sub-matrices of $\boldsymbol{F}^{-1}$ by superscripts, we have $\boldsymbol{F}_{\tau\omega}\boldsymbol{F}^{\omega\tau} = \boldsymbol{I}$ and $\boldsymbol{F}^{\omega\omega}\boldsymbol{F}_{\omega\tau} = \mathbf{0}$. Thus the updating function $\boldsymbol{u}(\boldsymbol{\xi})$ of (6) can be rewritten as follows

$$\boldsymbol{u}_{\omega}(\boldsymbol{\omega}) = \boldsymbol{\omega} + \boldsymbol{F}^{\omega\omega}\boldsymbol{e} + \boldsymbol{F}^{\omega\tau}h(\boldsymbol{\omega}), \quad \boldsymbol{u}_{\tau}(\boldsymbol{\omega}) = \boldsymbol{F}^{\tau\omega}\boldsymbol{e} + \boldsymbol{F}^{\tau\tau}h(\boldsymbol{\omega}),$$

neither of which is a function of $\boldsymbol{\tau}$. As the updating of the Lagrange multipliers does non depend on the estimation for $\boldsymbol{\tau}$ at previous step, the algorithm essentially searches in the space of $\boldsymbol{\omega}$. Hence, inserting a step size is only required for updating $\boldsymbol{\omega}$ and we propose, following Bergsma (1997) to use the following basic updating equations with an added step size, $0 < \text{step}^{(k)} \leq 1$:

$$\boldsymbol{\omega}^{(k+1)} = \boldsymbol{\omega}^{(k)} + \text{step}^{(k)}\{\boldsymbol{F}^{\omega\omega(k)}\boldsymbol{e}^{(k)} + \boldsymbol{F}^{\omega\tau(k)}h(\boldsymbol{\omega}^{(k)})\},$$

where $\boldsymbol{e}^{(k)} = \boldsymbol{n} - \hat{\boldsymbol{\mu}}^{(k)}$ and where $\boldsymbol{F}^{\omega\omega(k)}$ and $\boldsymbol{F}^{\omega\tau(k)}$ are two sections of $\hat{\boldsymbol{F}}^{-1}$ at cycle k. We chose the step size by a simple step halving criterion, but more sophisticated step size rules could also be considered. A discussion on the choice of the step size may be found in Bergsma (1997) and in Bergsma *et al.* (2009). Note that the algorithm's updates take place in the rectangular space $\mathbf{R}^t$ of $\boldsymbol{\omega}$ rather than the not necessarily rectangular space Λ of the marginal log-linear parameters which may not be variation independent. The algorithm converges if it is started from suitable initial estimates of $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$. While usually a zero vector is a good choice for $\boldsymbol{\tau}$, we found empirically that the number of iterations to convergence can be reduced substantially by using as a starting value for $\boldsymbol{\omega}$ an approximate maximum likelihood estimate based on results by Cox & Wermuth (1990) and Roddam (2004). At convergence, we obtain the maximum likelihood estimates $\hat{\boldsymbol{\mu}} = \exp(\hat{\boldsymbol{\omega}})$ and $\hat{\boldsymbol{\pi}} = N^{-1}\hat{\boldsymbol{\mu}}$ and the asymptotic covariance matrices

$$\mathrm{cov}(\hat{\boldsymbol{\omega}}) = \hat{\boldsymbol{F}}^{\omega\omega}, \qquad \mathrm{cov}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{H}_{sat}\hat{\boldsymbol{F}}^{\omega\omega}\boldsymbol{H}_{sat}^{\mathrm{T}}, \text{ with } \boldsymbol{H}_{sat} = \boldsymbol{D}_{\hat{\mu}}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{D}_{T\hat{\mu}}^{-1}\boldsymbol{C}^{\mathrm{T}}.$$

6 An application

In this section we illustrate the proposed disconnected set parameterization and the fitting algorithm for marginal independence models by using a real data set. It is rare that a pure marginal independence model is useful in isolation and thus usually it is interpreted in combination with other graphical models. However, the problem of simultaneous testing of multiple marginal independencies in a general contingency table is often present in applications and it can be carried out with the technique discussed in this paper. All the computations were programmed in the R language (R Development Core Team, 2008).

[Table 4 about here.]

The data are collected in a large contingency table including two ordinal variables with three levels. In the analysis these variables are treated as nominal variables using the baseline contrasts (2). Although the nature of the variables could be handled by using other more appropriate contrasts, as explained in Bartolucci *et al.* (2007), the fit of the marginal independence model is nevertheless invariant. table 4 summarizes observations for 13067 individuals on 6 variables obtained from as many questions taken from the U.S. General Social Survey (Davis *et al.*, 2007) during the years 1972-2006. The variables are reported below with the original name in the GSS Codebook:

- C CAPPUN: do you favor or oppose death penalty for persons convicted of murder? (1=favor, 2=oppose)
- F CONFINAN: confidence in banks and financial institutions (1= a great deal, 2= only some, 3= hardly any)
- G GUNLAW: would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun? (1=favor, 2=oppose)
- J SATJOB: how satisfied are you with the work you do? (1 = very satisfied, 2= moderately satisfied, 3 = a little dissatisfied, 4= very dissatisfied). Categories 3 and 4 of SATJOB were merged together.

S SEX: Gender (f,m)

A ABRAPE: do you think it should be possible for a pregnant woman to obtain legal abortion if she became pregnant as a result of rape? (1= yes, 2 = no)

[Figure 3 about here.]

In data sets of this kind there are a large number of missing values and the table used in this example collects only individuals with complete observations. Therefore, the following exploratory analysis is intended to be only an illustration with a realistic example. From a first analysis of the data, the following marginal independencies are not rejected by the chi-squared goodness of fit test statistic

$$\begin{array}{cccc} F \perp\!\!\!\perp CA & G \perp\!\!\!\perp JA & J \perp\!\!\!\perp GS & A \perp\!\!\!\perp FG \\ \chi^2_6 = 6.7 & \chi^2_5 = 3.3 & \chi^2_6 = 8.1 & \chi^2_5 = 2.1 \end{array}$$

and thus they suggest the independence model represented by the bi-directed graph in figure 3(a). Fitting this model, under the multinomial sampling assumption, we obtain an adequate fit with a deviance of 17.29 on 17 degrees of freedom. The modified Aitchison and Silvey's algorithm converges after 13 iterations. The encoded independencies cannot be represented by a directed acyclic graph model with the same observed variables, because the graph contains at least one subgraph which is a chordless 4-chain. The disconnected set parameterization defined by the ordered decomposable sequence

$$\mathcal{M}_G = \{CF, FA, GJ, GA, JS, CFA, FGA, GJS, GJA, CFGJSA\}$$

is variation independent. Alternatively, by searching in the class of graphical log-linear models with the backward stepwise selection procedure of MIM (Edwards, 2000) we found a model with a deviance of 103.16 over 110 degrees of freedom. The model graph is shown in figure 3(b). Other selection procedures show however that there are several equally well fitting models. The chosen undirected graph is slightly simpler (two fewer edges) than the bi-directed graph. As anticipated, the number of constraints on parameters is much higher. From the inspection of the studentized multivariate logistic estimates, we noticed that the higher order log-linear parameters are almost all not significant and thus we fitted a reduced model, by further restricting to zero all the log-linear parameters of order higher than two, obtaining a deviance of 108.34 on 118 degrees of freedom. The estimates of the remaining nonzero two-factor log-linear parameters are shown in table 5. These are estimated local log odds-ratios in the selected two-way marginal tables and they have the expected signs. By comparison, the fitted non-graphical log-linear model with the graph of figure 3(b), with additional zero constraints on the log-linear parameters of order higher than two, leads to a chi-squared goodness of fit of 118.49 on 119 degrees of freedom. Both models thus appear adequate.

[Table 5 about here.]

7 Discussion

The discrete models based on marginal log-linear models by Bergsma & Rudas (2002) form a large class that includes several discrete graphical models. The undirected graph

models and the chain graph models under the classical (Lauritzen, Wermuth, Frydenberg) interpretation can be parameterized as marginal log-linear models. For an introduction see Rudas *et al.* (2006). This paper shows that the discrete bi-directed graph models under the global Markov property are included in the same class by specifying the constraints appropriately. In general, three main criteria were considered in choosing a marginal log-linear parameterization.

(a) Upward compatibility: if the parameters have a meaning that is invariant across different marginal distributions, then the interpretations remain the same when a sub-model is chosen. We saw that the multivariate logistic parameterization has this property.

(b) Modelling considerations: the parameterization should contain all the parameters that are of interest for the problem at hand. For example, in a regression context where some variables are prior to others, effect parameters conditional on the explanatory variables are most meaningful. In the seemingly unrelated regression problem of example 4, the chosen parameters have the interpretation of logistic regression coefficients.

(c) Variation independence: if the parameter space is the whole Euclidean space, this has certain advantages. First, parameter interpretation is simpler, because in a certain sense different parameters measure different things. Second, in a Bayesian context, prior specification is easier. Finally, the problem of out-of-bound estimates (see Qaqish & Ivanova, 2006) when transforming parameters to probabilities is avoided. In the examples, we always found a variation independent parameterization, but a characterization of the class of bi-directed graphs admitting a variation independent complete and hierarchical marginal log-linear parameterization is an open problem.

The three criteria are in some cases conflicting: typically variation independence is obtained at the expense of upward compatibility.

The multivariate logistic parameterization has some analogies with the Möbius parameterization recently proposed by Drton & Richardson (2008) for binary marginal independence models, which is based on a minimal set of marginal probabilities identifying the joint distribution. These authors discuss the type of constraints on the Möbius parameters needed to specify a marginal independence, showing that they take a simple multiplicative form. The same constraints are defined by zero restrictions on marginal log-linear parameters in our approach. Even if the parametric space can be awkward, this problem is handled by a fitting algorithm that operates in the space of the expected frequencies, while the parameters are only used to define the independence constraints. Moreover, the definition of the models through the complete specification of the marginal log-linear parameters gives some advantage when there is a mixture of nominal and ordinal variables because it allows the definition of parameters for both types of variables using the theory of generalized marginal interactions by Bartolucci *et al.* (2007). This opens the way to defining subclasses of discrete graphical models specifying equality and inequality constraints.

The proposed algorithm for maximum likelihood fitting of the bi-directed graph model is a very general algorithm for constrained optimization based on Lagrange multipliers.

It is essentially a modification by Bergsma (1997) of the algorithm of Aitchison & Silvey (1958) and Lang & Agresti (1994). Related works on algorithms and generalizations have been proposed, for instance, by Molenberghs & Lesaffre (1994), Glonek & McCullagh (1995), Colombi & Forcina (2001) and Bergsma *et al.* (2009). An alternative algorithm could be obtained by extending to constrained optimization the multiparameter version of the algorithm by Jensen *et al.* (1991) which is potentially more stable and does not require complicated step size adjustment strategies.

The main advantage of Langrange multiplier based method is its generality (it can be applied to all models defined by equality and inequality constraints on the marginal log-linear parameters). As previously stated, the algorithm does not require further iterative procedures for computing, at each step, the inverse transformation from the marginal log-linear parameters to the cell probabilities. Thus, the risk of incompatible estimates that could arise from the lack of variation independence is avoided. Drawbacks are, as for many gradient-based algorithms, that convergence is not guaranteed and that for large tables, computation of the information matrix can become computationally rather heavy. However, empirical evidence shows that convergence is achieved in a relative small number of iterations by including a step adjustment. An alternative algorithm, with convergence guarantees, is the Iterative Conditional Fitting algorithm, proposed by Drton & Richardson (2008) for binary bi-directed graph models in the Möbius parameterization. A comparison between the two algorithms in terms of performance, speed and memory requirements needs further investigation.

Acknowledgement

We thank Nanny Wermuth for helpful discussions. The work of the first two authors was partially supported by MIUR, Rome, under the project PRIN 2005132307. The authors wish to thank the editor and two referees for their constructive comments and suggestions.

References

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–828.

Bartolucci, F., Colombi, R. & Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statist. Sinica* **17**, 691–711.

Bergsma, W., Croon, M. & Hagenaars, J. A. (2009). *Marginal models for dependent, clustered, and longitudinal categorical data.* Springer: New York.

Bergsma, W. P. (1997). *Marginal models for categorical data.* Ph.d thesis, University of Tilburg.

Bergsma, W. P. & Rudas, T. (2002). Marginal log-linear models for categorical data. *Ann. Statist.* **30**, 140 – 159.

Bertsekas, D. P. (1982). *Constrained optimization and Lagrange multiplier methods.* Academic Press, New York.

Colombi, R. & Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variabels. *Biometrika* **88**, 1007–1019.

Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* **8**, 204–218, 247–277.

Cox, R. D. & Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 747– 761.

Davis, J., Smith, T. & Marsden, J. A. (2007). *General Social Surveys Cumulative Codebook: 1972-2006*. NORC: Chicago.

Drton, M. (2008). Iterative conditional fitting for discrete chain graph models. In P. Brito, ed., *Compstat 2008 - Proceedings in Computational Statistics: 18th symposium*. Physica-Verlag, Heidelberg, pp. 93–104.

Drton, M. & Richardson, T. S. (2008). Binary models for marginal independence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 287–309.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Verlag, New York, (2nd ed.) edn.

Glonek, G. J. N. & McCullagh, P. (1995). Multivariate logistic models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 533–546.

Jensen, S., Johansen, S. & Lauritzen, S. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78(4)**, 867–877.

Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23**, 105–116.

Kauermann, G. (1997). A note on multivariate logistic models for contingency tables. *Aust. J. Statist.* **39**, 261–276.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.* **24**, 726–752.

Lang, J. B. & Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89(426)**, 625–632.

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

Lienert, G. A. (1970). Konfigurationsfrequenzalyse einiger Lysergsaurediathylamid-Wirkungen. *Arzneimittellorschung* **20**, 912–913.

Marchetti, G. M. & Lupparelli, M. (2008). Parameterization and fitting of a class of discrete graphical models. In P. Brito, ed., *Compstat 2008 - Proceedings in Computational Statistics: 18th symposium*. Physica-Verlag, Heidelberg, pp. 117–128.

Molenberghs, G. & Lesaffre, E. (1994). Marginal modelling of multivariate categorical data. *J. Amer. Statist. Assoc.* **89**, 633–644.

Qaqish, B. F. & Ivanova, A. (2006). Multivariate logistic models. *Biometrika* **93**, 1011–1017.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, T. S. (2003). Markov property for acyclic directed mixed graphs. *Scand. J. Statist.* **30**, 145–157.

Richardson, T. S. & Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30**, 962–103.

Roddam, A. W. (2004). An approximate maximum likelihood procedure for parameter estimation in multivariate discrete data regression models. *J. Appl. Stat.* **28**, 273– 279.

Rudas, T. & Bergsma, W. P. (2004). On applications of marginal models for categorical data. *Metron* **LXII**, 1– 25.

Rudas, T., Bergsma, W. P. & Németh, R. (2006). Parameterization and estimation of path models for categorical data. In A. Rizzi & M. Vichi, eds., *Compstat 2006 Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, pp. 383–394.

Wermuth, N. (1998). Pairwise independence. In P. Armitage & T. Colton, eds., *Encyclopedia of biostatistics*. Wiley, New York, pp. 3244–324.

Wermuth, N. & Cox, D. R. (1992). On the relation between interactions obtained with alternative codings of discrete variables. *Methodika* **VI**, 76–85.

Wermuth, N., Cox, D. R. & Marchetti, G. M. (2006). Covariance chains. *Bernoulli* **12**, 841–862.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley.

Giovanni M. Marchetti, Dipartimento di Statistica "G. Parenti", University of Florence
viale Morgagni, 59, 50134 Firenze, Italy.
E-mail: giovanni.marchetti@ds.unifi.it

List of Figures

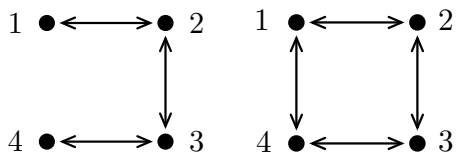

Figure 1: *Two bi-directed graphs.* (a) *The 4-chain with independencies* $4 \perp\!\!\!\perp 12$ *and* $1 \perp\!\!\!\perp 34$. (b) *The chordless 4-cycle with independencies* $1 \perp\!\!\!\perp 3$ *and* $2 \perp\!\!\!\perp 4$.

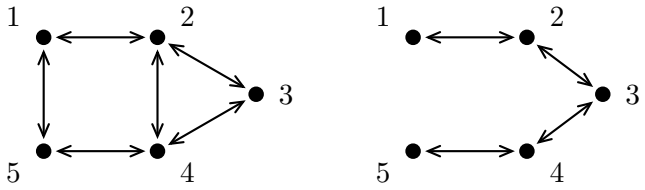

Figure 2: *Two bi-directed graphs in five nodes. The independencies implied by the connected set Markov property (or, equivalently, the global Markov property) are:* (a) $1 \perp\!\!\!\perp 34$, $3 \perp\!\!\!\perp 15$ *and* $5 \perp\!\!\!\perp 23$*;* (b) $1 \perp\!\!\!\perp 3 \perp\!\!\!\perp 5$, $1 \perp\!\!\!\perp 345$, $12 \perp\!\!\!\perp 45$ *and* $123 \perp\!\!\!\perp 5$.

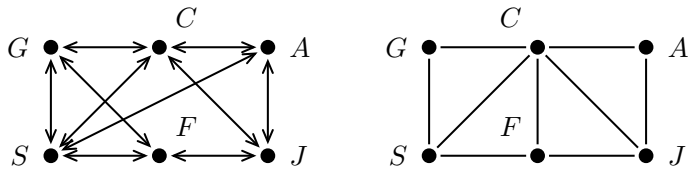

Figure 3: *Data from the U.S. General Social Survey 1972-2006. (a) A bi-directed graph model ($\chi^2_{17} = 17.29$). (b) A graphical log-linear model ($\chi^2_{110} = 103.16$).*

List of Tables

Table 1: *Data by Lienert (1970) concerning symptoms after LSD-intake. OR is the conditional odds-ratio between X_1 and X_2 given X_3. The frequencies show evidence of pairwise independence, but not mutual independence.*

	X_3	1		2	
X_1	X_2	1	2	1	2
1		21	5	4	16
2		2	13	11	1
OR		27.3		0.023	

Table 2: *Comparison between two parameterization of the discrete chordless 4-chain model of figure 1(a): ($\boldsymbol{\eta}$) with bi-directed edges; ($\boldsymbol{\theta}$) with undirected edges.*

Terms	1	2	3	4	12	13	14	23	24	34	123	124	134	234	1234
$\boldsymbol{\eta}$	$\boldsymbol{\eta}^{1}$	$\boldsymbol{\eta}^{2}$	$\boldsymbol{\eta}^{3}$	$\boldsymbol{\eta}^{4}$	$\boldsymbol{\eta}^{12}$	$\mathbf{0}$	$\mathbf{0}$	$\boldsymbol{\eta}^{23}$	$\mathbf{0}$	$\boldsymbol{\eta}^{34}$	$\boldsymbol{\eta}^{123}$	$\mathbf{0}$	$\mathbf{0}$	$\boldsymbol{\eta}^{234}$	$\boldsymbol{\eta}^{1234}$
$\boldsymbol{\theta}$	$\boldsymbol{\theta}_{1}$	$\boldsymbol{\theta}_{2}$	$\boldsymbol{\theta}_{3}$	$\boldsymbol{\theta}_{4}$	$\boldsymbol{\theta}_{12}$	$\mathbf{0}$	$\mathbf{0}$	$\boldsymbol{\theta}_{23}$	$\mathbf{0}$	$\boldsymbol{\theta}_{34}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$

Table 3: *Comparison of four parameterizations for the bi-directed graph model G of figure 1(a). One-factor log-linear parameters are omitted. The columns of zero constrained parameters have a boldface label.*

Terms	12	**13**	**14**	23	**24**	34	123	**124**	**134**	234	1234
$\boldsymbol{\eta}$	$\boldsymbol{\eta}^{12}$	$\boldsymbol{\eta}^{13}$	$\boldsymbol{\eta}^{14}$	$\boldsymbol{\eta}^{23}$	$\boldsymbol{\eta}^{24}$	$\boldsymbol{\eta}^{34}$	$\boldsymbol{\eta}^{123}$	$\boldsymbol{\eta}^{124}$	$\boldsymbol{\eta}^{134}$	$\boldsymbol{\eta}^{234}$	$\boldsymbol{\eta}^{1234}$
$\mathcal{M}_G$	$\boldsymbol{\lambda}_{12}^{124}$	$\boldsymbol{\lambda}_{13}^{13}$	$\boldsymbol{\lambda}_{14}^{14}$	$\boldsymbol{\lambda}_{23}^{1234}$	$\boldsymbol{\lambda}_{24}^{24}$	$\boldsymbol{\lambda}_{34}^{134}$	$\boldsymbol{\lambda}_{123}^{1234}$	$\boldsymbol{\lambda}_{124}^{124}$	$\boldsymbol{\lambda}_{134}^{134}$	$\boldsymbol{\lambda}_{234}^{1234}$	$\boldsymbol{\lambda}_{1234}^{1234}$
$\mathcal{M}'_G$	$\boldsymbol{\lambda}_{12}^{124}$	$\boldsymbol{\lambda}_{13}^{134}$	$\boldsymbol{\lambda}_{14}^{14}$	$\boldsymbol{\lambda}_{23}^{1234}$	$\boldsymbol{\lambda}_{24}^{124}$	$\boldsymbol{\lambda}_{34}^{134}$	$\boldsymbol{\lambda}_{123}^{1234}$	$\boldsymbol{\lambda}_{124}^{124}$	$\boldsymbol{\lambda}_{134}^{134}$	$\boldsymbol{\lambda}_{234}^{1234}$	$\boldsymbol{\lambda}_{1234}^{1234}$
$\mathcal{M}''_G$	$\boldsymbol{\lambda}_{12}^{124}$	$\boldsymbol{\lambda}_{13}^{134}$	$\boldsymbol{\lambda}_{14}^{134}$	$\boldsymbol{\lambda}_{23}^{1234}$	$\boldsymbol{\lambda}_{24}^{124}$	$\boldsymbol{\lambda}_{34}^{134}$	$\boldsymbol{\lambda}_{123}^{1234}$	$\boldsymbol{\lambda}_{124}^{124}$	$\boldsymbol{\lambda}_{134}^{134}$	$\boldsymbol{\lambda}_{234}^{1234}$	$\boldsymbol{\lambda}_{1234}^{1234}$

Table 4: *Data from U.S. General Social Survey.*

				F	1			2			3		
S	C	G	A	J	1	2	3	1	2	3	1	2	3
m	1	1	1		410	241	80	691	556	187	192	148	84
			2		71	31	9	109	64	34	27	26	15
		2	1		181	128	42	307	284	82	84	93	41
			2		41	17	5	61	35	20	18	13	5
	2	1	1		96	77	29	163	151	76	58	55	27
			2		34	18	7	58	36	15	17	13	6
		2	1		29	37	4	55	54	31	22	26	17
			2		16	6	6	16	16	7	10	7	2
f	1	1	1		552	353	145	899	793	265	180	162	94
			2		98	60	15	186	122	47	40	23	14
		2	1		133	74	33	219	164	66	36	47	24
			2		25	15	1	54	40	13	14	6	4
	2	1	1		228	153	60	356	343	166	95	80	41
			2		75	45	12	125	116	34	25	20	12
		2	1		41	25	13	64	56	22	15	14	11
			2		17	6	1	19	18	6	3	3	2

Table 5: *Estimates of two-factor log-linear parameters for the bi-directed graph model of figure 3(a) with additional zero restrictions on higher order terms. The column z contains the fitted studentized estimates.*

Margin	Parameter	Estimate	z	Margin	Parameter	Estimate	z
CG	(1)	−0.38	−7.86	FG	(1)	−0.01	−0.28
CJ	(1)	0.10	2.43		(2)	0.15	2.31
	(2)	0.25	4.37	FJ	(1)	0.29	6.73
CS	(1)	0.46	11.51		(2)	0.34	5.46
CA	(1)	0.56	11.40		(3)	0.34	5.43
GS	(1)	−0.77	−18.43		(4)	0.75	9.39
JA	(1)	−0.21	−4.18	FS	(1)	−0.004	−0.09
	(2)	−0.24	−3.43		(2)	−0.36	−6.43
SA	(1)	0.18	3.85				