arXiv:0707.3794v1 [math.ST] 25 Jul 2007

BINARY MODELS FOR MARGINAL INDEPENDENCE

MATHIAS DRTON AND THOMAS S. RICHARDSON

Abstract. Log-linear models are a classical tool for the analysis of contingency tables. In particular, the subclass of graphical log-linear models provides a general framework for modelling conditional independences. However, with the exception of special structures, marginal independence hypotheses cannot be accommodated by these traditional models. Focusing on binary variables, we present a model class that provides a framework for modelling marginal independences in contingency tables. The approach taken is graphical and draws on analogies to multivariate Gaussian models for marginal independence. For the graphical model representation we use bi-directed graphs, which are in the tradition of path diagrams. We show how the models can be parameterized in a simple fashion, and how maximum likelihood estimation can be performed using a version of the Iterated Conditional Fitting algorithm. Finally we consider combining these models with symmetry restrictions.

1. Introduction

In seminal work Anderson (1969, 1970, 1973) studied Gaussian models defined by hypotheses that are linear in covariances. Such hypotheses include as a special case, zero restrictions on covariance matrices. These restrictions correspond to marginal independences, which may arise for example through confounding effects of unobserved variables (Cox and Wermuth, 1993, 1996; Pearl and Wermuth, 1994; Richardson and Spirtes, 2002). For a graphical representation of zero restrictions on covariance matrices, Cox and Wermuth (1993, 1996) introduced *covariance graphs*: each variable is represented by a vertex; two vertices are linked by a dashed edge if the model does not set the corresponding covariance to zero. Dashed edges differentiate these graphs from undirected graphs, which represent zero hypotheses on the inverse covariance matrix (Lauritzen, 1996). More recently, a number of authors have used bi-directed edges ($\leftrightarrow$) in place of dashed edges which is consistent with Sewall Wright's (1921) path diagram notation; compare Figures 1, 2 and 4(a) below. Covariance graph models have appeared in several different contexts (e.g., Butte et al., 2000; Diaconis and Evans, 2002; Grzebyk et al., 2004; Mao et al., 2004). Maximum likelihood (ML) estimation and likelihood ratio (LR) tests in these Gaussian models can be carried out using the Iterative Conditional Fitting

Key words and phrases. bi-directed graph, covariance graph, graphical Markov model, iterative conditional fitting, maximum likelihood estimation, Möbius inversion.

algorithm (Drton and Richardson, 2003; Chaudhuri et al., 2007), which is implemented in the ‘ggm’ package in R (Marchetti, 2006).

There have been several efforts aimed at developing binary models with analogous independence structure. Kauermann (1997) uses the multivariate logistic (m-logit) transformation due to McCullagh (1989); McCullagh and Nelder (1989); Glonek and McCullagh (1995), which consists of selecting the highest order interaction term from every margin (see also Bergsma and Rudas, 2002b). Cox’s (1993) assumes that the joint distribution is quadratic exponential, and then approximates marginal distributions via series expansions. An alternative approach is to use the nonparametric concept of independence in order to form models for categorical data that are analogous to Gaussian models. Many existing discrete models, such as the popular graphical log-linear models for modelling conditional independence in contingency tables are often motivated this way (Wermuth, 1976; Darroch et al., 1980). In this paper we take this route to developing a general framework for modelling marginal independence that is a natural counterpart to graphical log-linear models.

For an example of a marginal independence pattern that cannot be represented using log-linear models but that our new models can accomodate very naturally suppose that we are investigating the relationship between alcohol dependence and depression. We have data from female mono-zygotic twins, indicating whether or not each twin is alcohol dependent (A_i) and whether or not they suffer from major depression (D_i); see Table 1. Consider the two graphs shown in Figure 1. Both hypothesize that for each twin there are independent factors relating to individual experiences (S_i) which influence both alcoholism and depression; however, graph (b) hypothesizes in addition that there is a single genetic factor which influences both traits, while graph (a) supposes that there is no such single factor, and that G_A, G_D, S_1 and S_2 are mutually independent. Graph (b) does not imply any independence restrictions relating the observed variables, while graph (a) implies that

$$A_1 \perp\!\!\!\perp D_2 \quad \text{and} \quad A_2 \perp\!\!\!\perp D_1; \tag{1}$$

using the notation of Dawid (1979). Under (a) one twin’s alcohol dependence status is independent of the other twin’s depression status. Note that we do not make any assumption concerning the marginal distributions of the unobserved variables. In particular, testing the hypothesis (1) provides a way of testing the scientific hypothesis leading to graph (a) without having to specify the number of levels of the possibly complex genetic factors. This focus on implied independences is in the spirit of the work on ancestral graphs (Richardson and Spirtes, 2002) and summary graphs (Cox and Wermuth, 1996). We remark that Ekholm et al. (2006a,b) recently fit latent class models to twin data including those in Table 1. The precise relationship between latent class models

TABLE 1. Data on $n = 597$ pairs of twins; adapted from Kendler et al. (1992).

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	288	80	92	51
	$A_2 = 1$	15	9	7	10
$A_1 = 1$	$A_2 = 0$	8	4	8	9
	$A_2 = 1$	3	2	4	7

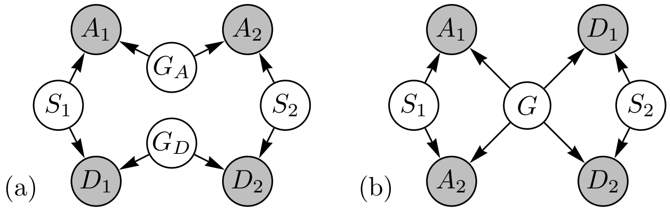

FIGURE 1. Possible generating models. Observed variables are shaded. (a) Separate genes relating to Alcohol (G_A) and Depression (G_D); (b) a common gene (G). S_j represents the personal experiences of twin j. Unobserved variables are hypothesized to be independent.

and the marginal independence models we discuss in the sequel is an open problem, but if two such models can be shown to coincide then the EM algorithm provides an alternative method for model fitting. However, in this context it should be noted that there exist Gaussian covariance graph models that cannot be parameterized by latent variable models (Richardson and Spirtes, 2002, §8.6).

If the variables were jointly Gaussian, then hypothesis (1) would restrict the appropriate two entries in the covariance matrix to zero. Hence, a likelihood ratio test of (1) could be performed by fitting the covariance matrix subject to this restriction. However, when the variables are binary, performing such a test is not at all straightforward. In particular, there does not exist a log-linear model that is equal to the family of binary distributions obeying (1). In fact, the marginal independence restrictions (1) correspond to complicated non-linear restrictions on the parameters of the log-linear expansion of the joint density of (A_1, A_2, D_1, D_2). The difficulty encountered here is an instance of the problem of lack of compatibility of margins in log-linear parametrizations (Glonek and McCullagh, 1995, p.534); see also McCullagh (1989). In this simple example, a practical solution might be to combine separate marginal tests, but there would be an obvious loss of efficiency in so doing. The methods developed in this paper

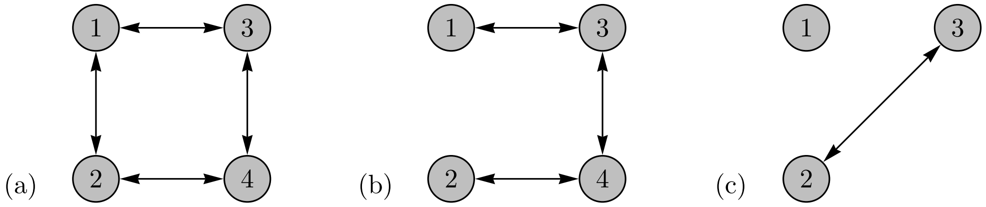

FIGURE 2. (a) a bi-directed four cycle; (b) a bi-directed four chain; (c) graph with two disconnected components.

allow the loss of efficiency to be avoided by providing models that capture precisely hypotheses like (1). The fitting algorithm we present allows tests that make use of all data available, such as LR- and χ^2-tests, to be performed.

The remainder of the paper is organized as follows. In §2 we describe the graphical representation of marginal independence patterns. This representation facilitates the understanding of marginal independence structures in multivariate normal distributions and provides the basis for our transfer of model structure to the binary case. This transfer yields models that are defined implicitly in terms of independence constraints. In §3, we show that a linear change of coordinates leads to a surprisingly simple characterization of marginal independence. This characterization immediately yields a multilinear model parameterization. ML estimation in the proposed models is discussed in §4 and the Iterative Conditional Fitting algorithm for computing ML estimates is developed in §5. In §6, the methodology is illustrated in an application to survey data. In the twin data example mentioned above, symmetry under permuting the labels 1 and 2 given to the twins is an interesting hypothesis. Combining such symmetry constraints with marginal independence is the topic of section §7. We conclude in §8, where connections to other work are discussed.

2. Bi-directed graphs and marginal independence

A bi-directed graph $G = (V, E)$ is a graph whose edges satisfy $(v, w) \in E$ if and only if $(w, v) \in E$. The edges are drawn bi-directed as $v \leftrightarrow w$ if $(v, w) \in E$, see Figure 2. Bi-directed graphs are special cases of the ancestral graphs considered in Richardson and Spirtes (2002) and the acyclic directed mixed graphs studied in Richardson (2003); see also Pearl (2000, p.146). If a vertex w is equal or adjacent to another vertex v in a bi-directed graph, then w is said to be a *spouse* of v, and we write $w \in \mathrm{Sp}(v)$. For a set $A \subseteq V$, we define $\mathrm{Sp}(A) = \cup(\mathrm{Sp}(v) \mid v \in A)$. Note that $A \subseteq \mathrm{Sp}(A)$ under this convention.

In graphical modelling, the Markov properties of a graph, i.e., independence statements associated with the graph, are used to define independence models for a random vector $X = (X_v \mid v \in V)$ whose index set is identified with the vertex set V of the graph. The independence models associated with bi-directed graphs are based on marginal independence, which is manifested in the *connected set Markov property* of Richardson (2003, §4). A vertex set $C \subseteq V$ is *connected* if every pair of vertices $v, w \in C$ are joined by a path on which every vertex is in C. The distribution of a random vector $X = (X_v \mid v \in V)$ is said to satisfy the connected set Markov property if

$$X_C \perp\!\!\!\perp X_{V \setminus \mathrm{Sp}(C)}, \tag{2}$$

whenever $\emptyset \neq C \subseteq V$ is a connected set. Algorithm E in Knuth (1968, p. 354) computes equivalence classes from a list of known equivalent pairs. This can be used to find the inclusion maximal connected sets in a given graph by letting the edges in the graph define the equivalent pairs.

A more exhaustive Markov property is the global Markov property, which requires all the marginal independences in (2), but also additional conditional independences. More precisely, the distribution of X satisfies the global Markov property of G if

$$A \text{ is separated from } B \text{ by } V \setminus (A \cup B \cup C) \text{ in } G \text{ implies } X_A \perp\!\!\!\perp X_B \mid X_C. \tag{3}$$

Here, A, B and C are disjoint subsets of V, and C may be empty. The separation in (3) is the usual graph-theoretic separation in which two sets $A, B \subset V$ are separated by a third set $D \subset V$ if any path from a vertex in A to a vertex in B contains a vertex in D. Despite the global Markov property being more exhaustive, a distribution satisfies the global Markov property if and only if it satisfies the connected set Markov property. Completeness of the global Markov property for bi-directed graphs follows from the completeness results for ancestral graphs (Richardson and Spirtes, 2002, Thm. 7.6). Note also the duality between (3) and the global Markov property for undirected graphs (Lauritzen, 1996, p. 32).

Example 1. (*Four-cycle*). The bi-directed graph depicted in Figure 2(a) represents the two pairwise independence relations:

$$X_1 \perp\!\!\!\perp X_4 \text{ and } X_2 \perp\!\!\!\perp X_3$$

under both the connected set, and global Markov properties. This graph represents the independence hypothesis considered in the introductory example; compare (1).

Example 2. (*Bi-directed four-chain*). Consider the bi-directed graph depicted in Figure 2(b). The connected set Markov property states

$$X_1 \perp\!\!\!\perp (X_2, X_4), \qquad X_2 \perp\!\!\!\perp (X_1, X_3), \qquad X_3 \perp\!\!\!\perp X_2, \qquad X_4 \perp\!\!\!\perp X_1.$$

The global Markov also states, for example, $X_1 \perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_2 \mid X_4$, and $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

Clearly, every singleton $\{v\}$ is a connected set and thus (2) requires that

$$X_v \perp\!\!\!\perp X_{V\setminus \mathrm{Sp}(v)}. \tag{4}$$

It follows that if the distribution of X satisfies the connected set Markov property, then it satisfies the pairwise Markov property which requires that $X_v \perp\!\!\!\perp X_w$ whenever $v \not\leftrightarrow w$. The converse is true for multivariate normal distributions (Kauermann, 1996, Prop. 2.2) but false in general. We note that the Markov property in (4) occurs in the combinatorial result known as the Lovász Local Lemma (Erdös and Lovász, 1975). In the next section we define models using the connected set (or equivalently the global) Markov property, and not the much less restrictive pairwise Markov property (see also Haber, 1986).

Example 3. (*Graph with two disconnected components*). The pairwise Markov property for the graph in Figure 2(c) requires $X_1 \perp\!\!\!\perp X_2$ and $X_1 \perp\!\!\!\perp X_3$, whereas the global and connected set Markov property also require the stronger condition that $X_1 \perp\!\!\!\perp (X_2, X_3)$. For example, consider the distribution of (X_1, X_2, X_3) given by

$$\begin{aligned} p_{000} &= 0.02, & p_{010} &= 0.03, & p_{100} &= 0.05, & p_{110} &= 0.10,\\ p_{001} &= 0.08, & p_{011} &= 0.12, & p_{101} &= 0.25, & p_{111} &= 0.35, \end{aligned}$$

where $p_{i_1 i_2 i_3} = P(X_1 = i_1, X_2 = i_2, X_3 = i_3)$. Then $X_1 \perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_3$ but $X_1 \not\!\perp\!\!\!\perp (X_2, X_3)$.

We conclude this discussion of Markov properties with a lemma that provides a useful characterization of joint distributions of discrete random vectors that obey the connected set Markov property. The lemma is based on the fact that every set $D \subseteq V$ that is not connected in G can be partitioned uniquely into inclusion-maximal connected sets $C_1, \dots, C_r$,

$$D = C_1 \dot\cup C_2 \dot\cup \cdots \dot\cup C_r. \tag{5}$$

Here, the symbol $\dot\cup$ denotes a union of disjoint sets.

Lemma 4. *Let $X = (X_v \mid v \in V)$ be a discrete random vector $X = (X_v \mid v \in V)$ taking values in the set $\mathcal{I}$. The joint distribution of X satisfies the connected set Markov property for a bi-directed graph $G = (V, E)$ if and only if for every disconnected set $D \subseteq V$ it holds that*

$$P(X_D = i_D) = P(X_{C_1} = i_{C_1}) P(X_{C_2} = i_{C_2}) \cdots P(X_{C_r} = i_{C_r}), \qquad \forall i \in \mathcal{I}, \tag{6}$$

where $C_1, \dots, C_r$ are the inclusion-maximal connected sets satisfying (5).

Proof. If P satisfies the connected set Markov property, then it also satisfies the global Markov property, from which we can deduce complete independence of the subvectors associated with the r connected sets in (5),

$$X_{C_1} \perp\!\!\!\perp X_{C_2} \perp\!\!\!\perp \dots \perp\!\!\!\perp X_{C_r}. \tag{7}$$

This complete independence clearly implies (6).

Conversely, let C be a connected set. Then $D = C \dot{\cup} (V \setminus \mathrm{Sp}(C))$ is a disconnected set, and (6) implies in particular $X_C \perp\!\!\!\perp X_{V \setminus \mathrm{Sp}(C)}$, which is (2). □

3. Binary marginal independence models

Let $X = (X_v \mid v \in V)$ be a random vector with binary components, i.e., X takes on values in the set $\mathcal{I} = \{0,1\}^V$, and let P be the joint distribution of X. (Note that to keep notation simple, we will often use the same letter to indicate both a set and its cardinality.) For $i = (i_v \mid v \in V) \in \mathcal{I}$, let

$$p_i = P(X_v = i_v \text{ for all } v \in V) \tag{8}$$

be the *joint cell probability of* i. The multivariate Bernoulli distribution of X is determined by the vector

$$p = \big(p_i \mid i \in \mathcal{I}\big) \tag{9}$$

in the $2^V - 1$ dimensional probability simplex Δ.

Using the Markov properties discussed in the previous section we can associate an independence model with a bi-directed graph $G = (V, E)$.

Definition 5. *The binary bi-directed graph model associated with G is defined as the family $\mathbf{B}(G)$ of probability distributions for a binary random vector $X = (X_v \mid v \in V)$ that obey the connected set Markov property (2) for G.*

We begin our study of the implicitly defined model $\mathbf{B}(G)$ by making a change of coordinates in the probability simplex. For $\emptyset \neq A \subseteq V$, we call

$$q_A = P(X_A = 0) = P(X_v = 0 \text{ for all } v \in A) \tag{10}$$

the *Möbius parameter associated with A*. If desired, q_A can be viewed as a moment for indicator variables associated with the designated levels of the considered binary variables, namely,

$$q_A = E\big(\textstyle\prod_{i \in A} 1_{\{X_i = 0\}}\big).$$

The $2^V - 1$ Möbius parameters can be computed from the joint cell probabilities p by the obvious summations

$$q_A = \sum_{i \in \mathcal{I} : i_A = 0} p_i, \tag{11}$$

where $i_A = (i_v \mid v \in A)$. The summations (11) define a map $\mu : \Delta \to \mathbb{R}^{2^V - 1}$ taking the vector of joint cell probabilities $p \in \Delta$ to the vector of Möbius parameters

$$q = \big(q_A \mid \emptyset \neq A \subseteq V\big). \tag{12}$$

We call the image $Q = \mu(\Delta)$ the *Möbius simplex*. This simplex has the 2^V vertices $t^{(A)}$, $A \subseteq V$, where for $\emptyset \neq B \subseteq V$ the B-th component of $t^{(A)}$ is equal to

$$t_B^{(A)} = \begin{cases} 1 & : \quad B \subseteq A, \\ 0 & : \quad B \not\subseteq A. \end{cases}$$

Clearly, $t^{(A)}$ is the image under μ of the distribution placing point mass on the cell $(0_A, 1_{V\setminus A})$.

Proposition 6. *The linear map*

$$\begin{array}{rccl} \mu : & \Delta & \to & Q \\ & p & \mapsto & q = (q_A \mid \emptyset \neq A \subseteq V) \end{array} \tag{13}$$

is bijective. Its inverse $\nu = \mu^{-1} : Q \to \Delta$ recovers the joint cell probabilities as alternating sums of Möbius parameters. Setting $q_\emptyset = 1$ we have

$$p_{0_A 1_{V\setminus A}} = P(X_A = 0, X_{V\setminus A} = 1) = \sum_{B : A \subseteq B} (-1)^{|B\setminus A|} q_B.$$

Proof. By definition of Q, the map μ is surjective. In order to verify injectivity and the claimed form of the inverse, define two functions Φ and Ψ on the power set of V. Let $\Phi(A) = q_{V\setminus A}$ for $A \subset V$ and $\Phi(V) = 1$, and $\Psi(A) = P(X_{V\setminus A} = 0, X_A = 1)$. Then $\Phi(A) = \sum_{B:B\subseteq A} \Psi(B)$ and the claim follows from the Möbius Inversion Lemma (Lauritzen, 1996, p.239). □

The maps ν and μ may be computed in $O(|V|2^{|V|-1})$ additions via the Fast Möbius Transform (Kennes and Smets, 1991). ADtrees (Moore and Lee, 1998) provide a memory-efficient data-structure for storing Möbius parameters. Moreover, the matrix for the map ν can be shown to have a Kronecker product structure; compare Jokinen (2006).

Example 7. (*Two binary random variables*). Consider two binary random variables, i.e., $V = \{1, 2\}$. Then the Möbius parameters are

$$q_1 = p_{00} + p_{01}, \quad q_2 = p_{00} + p_{10}, \quad q_{12} = p_{00}.$$

The joint cell probabilities can be recovered as

$$\begin{aligned} p_{00} &= q_{12}, & p_{01} &= q_1 - q_{12}, \\ p_{10} &= q_2 - q_{12}, & p_{11} &= 1 - q_1 - q_2 + q_{12}. \end{aligned}$$

The Möbius simplex is defined by the linear equalities expressing that p_i, written in terms of q, is in the unit interval $[0,1]$ for all $i \in \mathcal{I}$. In this example

$$Q = \big\{q = (q_1, q_2, q_{12}) \in [0,1]^3 \ : \ q_1 + q_2 - 1 \leq q_{12} \leq \min\{q_1, q_2\}\big\}$$

is a 3-dimensional simplex with vertices $(0,0,0)^t$, $(0,1,0)^t$, $(1,0,0)^t$, $(1,1,1)^t$.

As we show next, the constraints defining the independence model $\mathbf{B}(G)$ take on a simple form when expressed in terms of the Möbius parameter coordinates.

Theorem 8. *A probability vector $p \in \Delta$ belongs to the binary bi-directed graph model $\mathbf{B}(G)$ if and only if its Möbius parameters $q = \mu(p)$ satisfy that for every disconnected set $D \subseteq V$,*

$$q_D = q_{C_1} q_{C_2} \cdots q_{C_r}, \tag{14}$$

where $C_1, \ldots, C_r$ are the inclusion-maximal connected sets forming the partition (5).

Proof. By Lemma 4, $p \in \mathbf{B}(G)$ implies (14). Conversely, consider a vector $q \in Q$ satisfying (14), and let $p = \nu(q)$ be the associated probability vector. We show that $p \in \mathbf{B}(G)$ by verifying condition (6) in Lemma 4. We proceed by induction on the number of ones in the vector $i_D \in \{0,1\}^D$ appearing in (6), for some D, which we denote by $k \in \{0, 1, \ldots, V\}$.

By (14), the claim (14) holds for $k = 0$. Suppose that the claim holds for all $j < k$. Let v be such that $i_v = 1$ in i_D. Let $C_1, \ldots C_r$ be the partition of D into inclusion-maximal connected components, and suppose that $v \in C_\ell$. Then

$$\begin{aligned} P(X_D = i_D) &= P(X_{D\setminus\{v\}} = i_{D\setminus\{v\}}) - P(X_{D\setminus\{v\}} = i_{D\setminus\{v\}}, X_v = 0) \\ &= \big[P(X_{C_\ell\setminus\{v\}} = i_{C_\ell\setminus\{v\}}) - P(X_{C_\ell\setminus\{v\}} = i_{C_\ell\setminus\{v\}}, X_v = 0)\big] \\ &\qquad \times \prod_{j\neq\ell} P(X_{C_j} = i_{C_j}) \\ &= \prod_{j=1}^{r} P(X_{C_j} = i_{C_j}). \end{aligned}$$

The second equality follows from the induction hypothesis applied to $i_{D\setminus\{v\}}$, and to $\bar{\imath}_D = (i_{D\setminus\{v\}}, 0)$ since both vectors contain less than k ones. Hence, we have shown that (6) holds true for all disconnected sets $D \subseteq V$. □

Example 9. (*Four cycle*). For the bi-directed graph in Figure 2(a) we have 13 Möbius parameters associated with connected sets

$$q_1, q_2, q_3, q_4, \quad q_{12}, q_{13}, q_{24}, q_{34}, \quad q_{123}, q_{124}, q_{134}, q_{234}, \quad q_{1234}.$$

In order to define a distribution obeying $X_1 \perp\!\!\!\perp X_4$ and $X_2 \perp\!\!\!\perp X_3$, the Möbius parameters of the two disconnected sets must satisfy $q_{14} = q_1 q_4$ and $q_{23} = q_2 q_3$.

Theorem 8 can be read as providing a model parametrization. Let $Q_G = \mu(\mathbf{B}(G))$ be the Möbius parameter vectors defining a distribution in $\mathbf{B}(G)$. Let $\mathcal{C}(G)$ be the family of non-empty connected sets of G. Define T_G to be the set of vectors $(q_C \mid C \in \mathcal{C}(G)) \in \mathbb{R}^{\mathcal{C}(G)}$ of Möbius parameters of connected sets for which there exists a vector $\bar{q} \in Q_G$ with $\bar{q}_C = q_C$ for all $C \in \mathcal{C}(G)$.

Corollary 10. *Let $\nu_G : T_G \to \mathbf{B}(G)$ be the multilinear map defined by setting Möbius parameters of disconnected sets equal to the expression in (14), obtaining a vector $q \in \mathbb{R}^{2^V-1}$, and setting $p = \nu(q) \in \mathbf{B}(G)$. Then ν_G is a bijection, and we call it the Möbius parametrization of the model $\mathbf{B}(G)$.*

Since the Möbius parameters are related via inequalities (compare Example 7), this parametrization is not variation independent, but nevertheless is useful. The definition of the Möbius parameters is clearly not symmetric under re-labelling of the two states taken by the random variables. However, such re-labelling does not change the model $\mathbf{B}(G)$ because it is defined purely in terms of independence relations.

Corollary 11. *The dimension of the model $\mathbf{B}(G)$ equals $\dim(\mathbf{B}(G)) = |\mathcal{C}(G)|$, the number of non-empty connected sets in G.*

In contrast, the dimension of the (binary) graphical log-linear model based on the undirected graph with the same edges as G would be equal to the number of non-empty complete sets in G. Here a set $A \subseteq V$ is complete if any two vertices in A are adjacent. Since every complete set is connected, the dimension of the model $\mathbf{B}(G)$ is always larger than or equal to the dimension of the corresponding graphical log-linear model; compare Figure 3.

Corollary 12. *The family*

$$\mathbf{B}_+(G) = \{p \in \mathbf{B}(G) \,:\, p_i > 0 \ \textit{ for all } i \in \mathcal{I}\}$$

of distributions with positive joint cell probabilities in the binary bi-directed graph model $\mathbf{B}(G)$ forms a $|\mathcal{C}(G)|$-dimensional curved exponential family.

Proof. More precisely stated, we claim that $\mathbf{B}_+(G)$ is a $|\mathcal{C}(G)|$-dimensional smooth manifold in the natural parameter space of the exponential family formed by the interior of the probability simplex Δ^o. Let Q^o be the interior of the Möbius simplex Q, and Q^o_G the set of vectors in Q^o that satisfy the constraints (14) in Theorem 8. Let $d = 2^V - |\mathcal{C}(G)| - 1$ be the number of non-empty disconnected sets of G. Define the map $h : Q^o \to \mathbb{R}^d$ with coordinate functions $h_D(q) = q_D - q_{C_1}q_{C_2}\cdots q_{C_r}$, where $C_1, C_2, \ldots, C_r$ form the inclusion-maximal connected set partition of the non-empty disconnected set D; compare (5). Since h is C^∞, it is clear that $Q^o_G = h^{-1}(0)$ is a $|\mathcal{C}(G)|$-dimensional smooth manifold in $\mathbb{R}^{2^V-1}$; compare e.g. Thm. 1 in Geiger et al. (2001). Our claim is

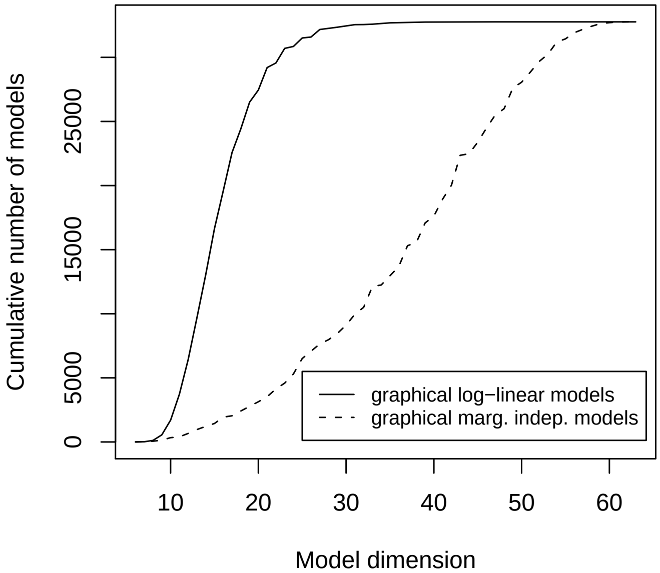

FIGURE 3. Cumulative number of models per dimension for $|V| = 6$ binary variables.

now established because the diffeomorphism ν maps Q^o to Δ^o, and it is well-known that there is a diffeomorphism between Δ^o (mean parameters) and the log-linear parameters (natural parameters of the exponential family). □

Remark 13. Instead of using the Möbius parameters in Theorem 8, we could have employed the *dependence ratios*

$$\tau_A = \frac{q_A}{\prod_{i\in A} q_i}$$

introduced by Ekholm et al. (1995); see also Ekholm et al. (2000, 2003), and Darroch and Speed (1983) where such ratios occur in specifying models termed *Lancaster additive*. The ratio τ_A compares the probability q_A computed from the joint distribution p to the corresponding probability under the complete independence distribution that has the same univariate marginals as p. Clearly, Theorem 8 also holds if we replace each Möbius parameter by the corresponding dependence ratio.

4. Maximum likelihood estimation

Assume we observe a sample of size n drawn from a distribution p in the binary bi-directed graph model $\mathbf{B}(G)$, giving rise to multinomially distributed counts $N(i)$, $i \in \mathcal{I}$. (For the link to Poisson sampling see Lauritzen, 1996, §4.2.1.) The probability of observing the particular counts $n(i) \in \mathbb{N}_0$, $i \in \mathcal{I}$, is equal to

$$P(N(i) = n(i),\ i \in \mathcal{I}) = \frac{n!}{\prod_{i\in\mathcal{I}} n(i)!} \prod_{i\in\mathcal{I}} p_i^{n(i)}, \tag{15}$$

where we set $0^0 := 1$. Hence, the likelihood function for the model $\mathbf{B}(G)$ is the map

$$\begin{aligned} L : \mathbf{B}(G) &\to \mathbb{R}, \\ p &\mapsto \frac{n!}{\prod_{i\in\mathcal{I}} n(i)!} \prod_{i\in\mathcal{I}} p_i^{n(i)}. \end{aligned} \tag{16}$$

Proposition 14. *An MLE of $p \in \mathbf{B}(G)$ always exists.*

Proof. As a subset of the probability simplex Δ, the model $\mathbf{B}(G)$ is bounded. It is also closed, hence compact, which in conjunction with the continuity of the likelihood function implies the claim. Closedness follows from the fact that if for two sets $A, B \subseteq V$, $X_A \perp\!\!\!\perp X_B$ under a sequence of probability distribution P_n with vector of joint cell probabilities $p_n \in \Delta$, then under a probability distribution P corresponding to a limit point $p \in \Delta$ of the sequence (p_n) it is also true that $X_A \perp\!\!\!\perp X_B$; compare Lauritzen (1996, Prop. 3.12). □

If all counts $n(i)$, $i \in \mathcal{I}$, are positive, then an MLE of $p \in \mathbf{B}(G)$ will actually have positive joint cell probabilities, i.e., lie in $\mathbf{B}_+(G)$. An open question is when an MLE exists in $\mathbf{B}_+(G)$ if some of the counts $n(i)$ are zero. For recent work on the analogous question in the case of hierarchical log-linear models see Eriksson et al. (2006). Another open problem concerns uniqueness of the MLE, i.e., can one find a graph G and (non-degenerate) counts $n(i)$ such that the likelihood function of $\mathbf{B}(G)$ has more than one local maximum?

Ignoring an additive constant the log-likelihood function for the model $\mathbf{B}(G)$ is of the form $\ell(p) = \sum_{i\in\mathcal{I}} n(i) \log p_i$. Using Proposition 6, we can express the log-likelihood function also in terms of Möbius parameters as

$$\begin{aligned} \ell : Q_G &\to \mathbb{R}, \\ q &\mapsto \sum_{A\subseteq V} n(0_A, 1_{V\setminus A}) \log\left[\sum_{B: A\subseteq B} (-1)^{|B\setminus A|} q_B\right], \end{aligned}$$

where $q_\emptyset = 1$. Further, $\ell(q)$ can be written in terms of the connected set Möbius parameters $(q_C \mid C \in \mathcal{C}(G))$ by replacing q_B for a disconnected set B by the appropriate product of connected set Möbius parameters; see (14).

For two subsets $A, W \subseteq V$, nested as $A \subseteq W$, define

$$p_A^W = P(X_A = 0, X_{W\setminus A} = 1).$$

In particular, if $W = V$, then $p_A^V = P(X_A = 0, X_{V\setminus A} = 1)$ is a joint cell probability. Similarly define n_A^V to be the frequency of observations in which $X_A = 0$, and $X_{V\setminus A} = 1$. Then the likelihood equations associated with the model $\mathbf{B}(G)$ are

$$\frac{\partial \ell}{\partial q_C} = \sum_{A:\mathrm{Sp}(C)\cap(A\setminus C)=\emptyset} (-1)^{|C\setminus A|} \frac{n_A^V}{p_A^V} p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)} = 0$$

for every (non-empty) connected set C in G. We prove this in the Appendix (Corollary 12), where we also compute the second derivative of $\ell(q)$, which yields the Fisher-information for $\mathbf{B}(G)$.

Having written the log-likelihood function as a function of the parameters $(q_C \mid C \in \mathcal{C}(G))$, it can be maximized using gradient-based ascent methods (see also Lang and Agresti, 1994; Bergsma and Rapcsák, 2005). We implemented such a method in the statistical programming environment R (R Development Core Team, 2004) using the routine 'nlm'. In doing this we found it beneficial to work with the logarithms of the parameters q_C because this linearizes (14); the examples we considered involve positive counts such that we may assume that q_C is positive and $\log q_C$ well-defined. In our experience, this approach works well for smaller and sparser graphs that induce a lower-dimensional model. However, for larger and denser graphs, such as in Figure 4(a), we found an alternative approach that focuses on the model-defining constraints to perform better. This alternative method, described in the next section, is the binary analogue to the Iterative Conditional Fitting (ICF) algorithm that was developed for ML fitting of Gaussian marginal independence models (Drton and Richardson, 2003; Chaudhuri et al., 2007). Binary ICF plays a role dual to the Iterative Proportional Fitting (IPF) algorithm used to fit hierarchical log-linear models.

5. Iterative conditional fitting

Starting from some feasible estimate in $\mathbf{B}(G)$, such as the uniform distribution, the ICF algorithm improves a current feasible estimate by cycling through the vertex set V and performing an update step for each one of the vertices. At the update step for variable $v \in V$ the marginal distribution $P^{X_{-v}}$ of the variables $-v = V \setminus \{v\}$ is fixed, and the conditional distribution $P^{X_v|X_{-v}}$ required to determine the joint distribution of $(X_v \mid v \in V)$ is estimated. This estimation is done subject to constraints that ensure that the newly determined joint distribution remains in the model $\mathbf{B}(G)$. In this presentation of ICF we assume that all observed counts $n(i)$, $i \in \mathcal{I}$, are positive, which in particular entails that they were drawn from a distribution $p \in \mathbf{B}_+(G)$. Moreover, maximizing the likelihood function over $\mathbf{B}(G)$ is equivalent to maximizing it over the submodel $\mathbf{B}_+(G)$, and we can assume that all joint distributions P considered in the sequel have positive joint cell probabilities $p_i > 0$. In the case of zero counts, which will be considered in future work, the conditional likelihood function considered in Algorithm 5.1 is still concave but need no longer be strictly concave. Hence, the possibility of optima on the boundary has to be taken into account.

For fixed marginal probability $P(X_{-v} = i_{-v})$, the joint cell probability $P(X_v = i_v, X_{-v} = i_{-v}) = 0$, $i \in \mathcal{I}$, is determined by the conditional parameter

$$\theta_v(i_{-v}) = \theta_v(X_{-v} = i_{-v}) := P(X_v = 0 \mid X_{-v} = i_{-v}).$$

Let $\mathcal{I}_{-v} = \{0,1\}^{V-1}$. Then there are $|\mathcal{I}_{-v}| = 2^{V-1}$ many parameters $\theta_v(i_{-v})$. Notice that if $v \in D$ then

$$\begin{aligned} P(X_D = 0) &= \sum_{B:D\subseteq B} P(X_B = 0, X_{V\setminus B} = 1) \\ &= \sum_{i_{-v}=(0_{B\setminus\{v\}},1_{V\setminus B})\in\mathcal{I}_{-v}\,:\,D\subseteq B} \theta_v(i_{-v})P(X_{-v} = i_{-v}). \qquad (17) \end{aligned}$$

In general, the binary bi-directed graph model $\mathbf{B}(G)$ imposes constraints on the conditional distribution $P^{X_v|X_{-v}}$. In order to specify the constraints in a non-redundant way, we focus on constraints of the form (14), rather than the equivalent conditional independence restrictions. Specifically, suppose that D is a disconnected set and that C is the inclusion-maximal connected subset of D containing v. By equation (14) we require

$$P(X_D = 0) = P(X_C = 0)P(X_{D\setminus C} = 0). \qquad (18)$$

Note that $D \setminus C$ may not be connected, so the model may require further factorization of $P(X_{D\setminus C} = 0)$. However, this only imposes a constraint on the fixed $P(X_{-v})$ margin, and so does not concern us here. We now express the constraint (18) as

$$\begin{aligned} &P(X_v = 0 \mid X_{D\setminus\{v\}} = 0)P(X_{D\setminus\{v\}} = 0) \\ &\quad = P(X_v = 0 \mid X_{C\setminus\{v\}} = 0)P(X_{C\setminus\{v\}} = 0)P(X_{D\setminus C} = 0). \qquad (19) \end{aligned}$$

(It is implicit here that if $C \setminus \{v\} = \emptyset$ then the second term on the right hand side is omitted.) Observe that only the first terms on each side depend on $\theta_v(\cdot)$. Using (17), the first term on the left hand side of (19) may be expressed as

$$\begin{aligned} &P(X_v = 0 \mid X_{D\setminus\{v\}} = 0) \\ &\quad = \sum_{j\in\{0,1\}^{V\setminus D}} P(X_v = 0, X_{V\setminus D} = j \mid X_{D\setminus\{v\}} = 0). \\ &\quad = \sum_{j\in\{0,1\}^{V\setminus D}} \theta_v(X_{V\setminus D} = j, X_{D\setminus\{v\}} = 0)P(X_{V\setminus D} = j \mid X_{D\setminus\{v\}} = 0). \qquad (20) \end{aligned}$$

Similarly, the first term on the right hand side of (19) may be expressed as

$$\begin{aligned} &P(X_v = 0 \mid X_{C\setminus\{v\}} = 0) \\ &\quad= \sum_{j\in\{0,1\}^{V\setminus C}} \theta_v(X_{V\setminus C} = j, X_{C\setminus\{v\}} = 0) P(X_{V\setminus C} = j \mid X_{C\setminus\{v\}} = 0). \end{aligned} \tag{21}$$

Now, if the set $D \setminus \{v\}$ was connected, then C and $D \setminus C$ would also be connected, contrary to the assumption. Since in the ICF algorithm we assume that all constraints on the marginal distribution of $X_{V\setminus\{v\}}$ hold, it follows that

$$P(X_{D\setminus\{v\}} = 0) = P(X_{D\setminus C} = 0) P(X_{C\setminus\{v\}} = 0).$$

(Again, both $C \setminus \{v\}$ and $D \setminus C$ may not be connected, so these terms may factorize further.) Since these terms are non-zero, they cancel from both sides of (19), leaving the constraint

$$\begin{aligned} &\sum_{j\in\{0,1\}^{V\setminus D}} \theta_v(j, 0_{D\setminus\{v\}}) P(X_{V\setminus D} = j \mid X_{D\setminus\{v\}} = 0) = \\ &\qquad\qquad \sum_{j\in\{0,1\}^{V\setminus C}} \theta_v(j, 0_{C\setminus\{v\}}) P(X_{V\setminus C} = j \mid X_{C\setminus\{v\}} = 0). \end{aligned} \tag{22}$$

It is important to note that for fixed margin $P^{X_{-v}}$ the constraints (22) are linear in the conditional parameters θ_v. The full set of constraints on the θ_v parameters may be obtained by considering every disconnected set D containing v and identifying the inclusion-maximal connected set $C \subset D$ containing v.

Let

$$\mathfrak{D}_v = \{D : D \subseteq V,\ v \in D,\ D \text{ is disconnected}\}.$$

For each set $D \in \mathfrak{D}_v$, we define $C_v(D)$ to be the inclusion-maximal connected subset of D containing v. The disconnected sets $\mathfrak{D}_v$ and the connected sets $C_v(D)$ can be computed in preprocessing. Then the ICF update for vertex v can be implemented as follows.

Algorithm 15. *Update step in Iterative Conditional Fitting.*
Input: A probability vector $p \in \mathbf{B}(G)$ and vertex v.
Output: A probability vector $\bar{p} \in \mathbf{B}(G)$ such that $L(\bar{p}) \geq L(p)$.
Step 1. Construct the $\mathfrak{D}_v \times \mathcal{I}_{-v}$ constraint matrix $A = (a_{rs})$, where for each pair $(D_r, j_s) \in \mathfrak{D}_v \times \mathcal{I}_{-v}$ we set

$$\begin{aligned} a_{rs} = {} & P(X_{V\setminus D_r} = (j_s)_{V\setminus D_r} \mid X_{D_r\setminus\{v\}} = 0)\, \mathrm{I}\{(j_s)_{D_r\setminus\{v\}} = 0\} \\ & - P(X_{V\setminus C_v(D_r)} = (j_s)_{V\setminus C_v(D_r)} \mid X_{C_v(D_r)\setminus\{v\}} = 0)\, \mathrm{I}\{(j_s)_{C_v(D_r)\setminus\{v\}} = 0\}. \end{aligned}$$

Here all probabilities are computed under the distribution induced by the probability vector p, and $\mathrm{I}(\cdot)$ is the indicator function.

Step 2. Maximize the strictly concave conditional log-likelihood function

$$\sum_{i_{-v}\in\mathcal{I}_{-v}} n(i_{-v},0)\log\theta(i_{-v}) + n(i_{-v},1)\log\{1-\theta(i_{-v})\}$$

subject to the linear constraints $A\theta = 0$, where $\theta = (\theta(i_{-v}) \mid i_{-v}\in\mathcal{I}_{-v})$ is the vector of all conditional parameters. (If all counts are positive, the inequality constraints $\theta\in[0,1]^{\mathcal{I}_{-v}}$ need not be considered explicitly.)

Step 3. Use the solution θ_v from step 2 to compute the new probability vector $\bar{p}\in\mathbf{B}(G)$ via

$$\bar{p}_{i_v i_{-v}} = \bar{P}(X_v = i_v, X_{-v} = i_{-v}) = \begin{cases} \theta_v(i_v)\, P(X_{-v} = i_v) & \text{if } i_v = 0,\\ [1-\theta_v(i_v)]\, P(X_{-v} = i_{-v}) & \text{if } i_v = 1. \end{cases}$$

The optimization problem in step 2 of the ICF update algorithm has a unique local maximum and is not difficult to solve. For example, one can employ the gradient projection method (Bertsekas, 1999, §2.3), which performs a line search along the direction of the gradient projected on the kernel of A. A line search based on the Armijo-rule ensures convergence of the gradient projection method. The computation of the $2^{V-1}\times 2^{V-1}$ projection matrix $I - A'(AA')^{-1}A$ requires the inversion of the $\mathfrak{D}_v\times\mathfrak{D}_v$ matrix AA' which is of full rank. However, the projection matrix has to be computed only once in order to solve the optimization problem in step (4) of Algorithm 15. Since the Hessian of the conditional log-likelihood function maximized in step 2 of Algorithm 15 is diagonal it is also feasible to employ second derivative information in a projected Newton method, in which θ is scaled by the matrix with diagonal elements equal to one over the square root of the diagonal elements of the Hessian. Since the Hessian depends on θ, the projection matrix in a projected Newton method has to be recomputed every time θ is updated. However, based on our experience with our implementation of ICF in R, employing the Hessian information is beneficial.

Having tackled the individual ICF updates we can run ICF from a feasible starting value. The algorithm then produces a sequence of feasible estimates whose accumulation points are solutions to the likelihood equations. In fact, the sequence is guaranteed to converge if there exist only finitely many solutions to the likelihood equations. These convergence guarantees follow from general results about iterative partial maximization algorithms (Drton and Eichler, 2006, Appendix), of which ICF is an incarnation. Here, 'partial maximization' refers to the fact that the update step for vertex v in Algorithm 15 maximizes the log-likelihood function $\ell(q)$, $q = (q_C \mid C\in\mathcal{C}(G))$ partially, namely when varying only components q_C for which C is a connected set containing vertex

v. Components q_C for connected sets not containing v remain fixed at the current estimates.

In the above we proceeded vertex-by-vertex and estimated the univariate conditional distribution of X_v given X_{-v}. In the Gaussian case, Chaudhuri et al. (2007) describe how to run the ICF algorithm with multivariate updates. In this variant, one chooses complete vertex sets $C \subseteq V$ and estimates, for fixed margin of $X_{-C} = X_{V\setminus C}$, the (multivariate) conditional distribution of X_C given X_{-C} under the marginal independence constraints. Such multivariate updates are also possible in the binary case discussed here. Let

$$\theta_C(i_{-C}) = \theta_C(X_{-C} = i_{-C}) := P(X_C = 0 \mid X_{-C} = i_{-C}).$$

Let $\mathcal{I}_{-C} = \{0,1\}^{V-|C|}$. Then (17) becomes

$$P(X_D = 0) = \sum_{i_{-C}=(0_{B\setminus\{C\}},1_{V\setminus B})\in\mathcal{I}_{-C}\,:\,D\subseteq B} \theta_C(i_{-C})P(X_{-C} = i_{-C}). \tag{23}$$

Since the set C is complete there are no equality constraints among the Möbius parameters q_A, $\emptyset \neq A \subseteq C$, and one can proceed similarly as in the discussion leading up to (22) to devise an analog to Algorithm 15 with multivariate updates over complete sets.

6. Example: Social survey data

Sociologists and political theorists have long been interested in the relationship between trust in social institutions and trust in other members of society (Putnam, 2002; Sztompka, 2000; Levi, 1998). Here, as an illustration of an exploratory analysis using binary independence models we examine seven questions relating to trust that are taken from the U.S. General Social Survey during the years 1975-94:

- Trust

  Generally speaking, would you say that most people can be trusted or that you can't be too careful in life. (Can Trust; Cannot Trust; Depends)
- Helpful

  Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves? (Helpful; Lookout for Self; Depends)
- Confidence in institutions

  I am going to name some institutions in this country. As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them? (A Great Deal; Only Some; Hardly Any)

  ConClerg: Organized religion
  ConLegis: Congress
  ConBus: Major Companies

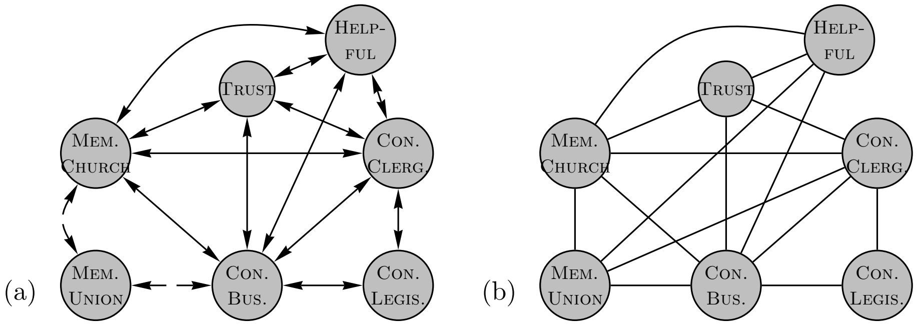

Figure 4. Analysis of Trust data. (a) Marginal independence model; dashed edges correspond to pairwise odds ratios less than one. (b) Classical graphical log-linear model.

- Membership of organizations

  Here is a list of various organizations. Could you tell me whether or not you are a member of each type? (Yes; No)

  MemUnion: Labour unions MemChurch: Churches

There were $13,486$ individuals who gave valid responses to all of these questions. For the purposes of illustration, for the questions relating to confidence in institutions we combine 'Some' and 'Hardly Any' to form a 'No' response; similarly for Trust we combine 'Cannot trust' with 'Depends' to form a 'No' group, and for Helpful we combine 'Take advantage' with 'Depends' to form a 'No' group. The counts are displayed in Table 2.

Using ICF in a backward stepwise selection we found the graph shown in Figure 4(a). Assuming that the data in Table 2 arose in multinomial sampling, we obtain a deviance of 32.67 over 26 degrees of freedom, when compared to the saturated model of no independence. Using an asymptotic χ^2-approximation a p-value of 0.172 is obtained and the model is found not to be contradicted by the data. Since some expected cell counts are small, the asymptotic approximation should be treated with some caution. In the selected model all variables are marginally associated with confidence in business, but it is interesting that confidence in congress is marginally associated only to the two other confidence variables. Similarly, union membership is not marginally associated with additional variables other than church membership and confidence in business; the graph implies

$$\textsc{ConLegis} \perp\!\!\!\perp \textsc{Helpful}, \textsc{Trust}, \textsc{MemUnion}, \textsc{MemChurch}, \quad \text{and}$$
$$\textsc{MemUnion} \perp\!\!\!\perp \textsc{Helpful}, \textsc{Trust}, \textsc{ConLegis}, \textsc{ConClerg}.$$

It is perhaps of little surprise that in the fitted distribution the marginal odds ratio between MemUnion and ConBus is less than one; it equals 0.83. Except for the odds

Table 2. Data from the U.S. General Social Survey relating to Trust.

					Helpful			
					Yes	Yes	No	No
Con.	Con.	Con.	Mem.	Mem.	Trust		Trust	
Bus.	Clerg.	Legis.	Church	Union	Yes	No	Yes	No
Yes	Yes	Yes	Yes	Yes	18	4	5	5
Yes	Yes	Yes	Yes	No	79	47	17	30
Yes	Yes	Yes	No	Yes	8	9	1	15
Yes	Yes	Yes	No	No	88	55	22	79
Yes	Yes	No	Yes	Yes	22	11	10	13
Yes	Yes	No	Yes	No	194	95	33	77
Yes	Yes	No	No	Yes	31	10	13	23
Yes	Yes	No	No	No	179	82	58	122
Yes	No	Yes	Yes	Yes	7	5	1	3
Yes	No	Yes	Yes	No	40	27	11	23
Yes	No	Yes	No	Yes	9	10	1	12
Yes	No	Yes	No	No	68	56	33	73
Yes	No	No	Yes	Yes	15	13	6	14
Yes	No	No	Yes	No	188	117	52	100
Yes	No	No	No	Yes	32	28	22	35
Yes	No	No	No	No	366	185	120	312
No	Yes	Yes	Yes	Yes	7	5	2	6
No	Yes	Yes	Yes	No	62	32	11	48
No	Yes	Yes	No	Yes	5	9	2	12
No	Yes	Yes	No	No	38	37	11	64
No	Yes	No	Yes	Yes	40	26	17	34
No	Yes	No	Yes	No	270	187	73	281
No	Yes	No	No	Yes	25	33	11	50
No	Yes	No	No	No	202	216	84	356
No	No	Yes	Yes	Yes	5	2	3	11
No	No	Yes	Yes	No	51	32	17	59
No	No	Yes	No	Yes	15	18	7	33
No	No	Yes	No	No	104	79	40	172
No	No	No	Yes	Yes	74	62	27	108
No	No	No	Yes	No	603	469	177	654
No	No	No	No	Yes	199	181	84	305
No	No	No	No	No	1002	920	460	1818

TABLE 3. ML estimate of the joint distribution under the symmetry group $\mathcal{S}_{\text{twin}}$; empirical distribution in parenthesis.

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	0.4824 (0.4824)	0.1441 (0.1340)	0.1441 (0.1541)	0.0854 (0.0854)
	$A_2 = 1$	0.0193 (0.0251)	0.0142 (0.0151)	0.0092 (0.0117)	0.0159 (0.0168)
$A_1 = 1$	$A_2 = 0$	0.0193 (0.0134)	0.0092 (0.0067)	0.0142 (0.0134)	0.0159 (0.0151)
	$A_2 = 1$	0.0050 (0.0050)	0.0050 (0.0034)	0.0050 (0.0067)	0.0117 (0.0117)

ratio between MemUnion and MemChurch, which is equal to 0.85, all other fitted pairwise odds ratios are greater than or equal to 1.

For purposes of comparison, in Figure 4(b) we include a classical graphical log-linear model obtained using the MIM program (Edwards, 2000) by backward stepwise selection, among all undirected models. This model has a deviance of 87.62 over 88 degrees of freedom. When comparing the undirected and bi-directed models it is quite striking that the undirected model contains one more edge, yet 62 fewer parameters. Observe that in the undirected graph union membership is also adjacent to the variables relating to confidence in clergy and whether or not people are helpful. We remark that latent variable models could be used for further analyses of these data.

7. Independence and symmetry

In this section we demonstrate how symmetry can be incorporated in the marginal independence models proposed earlier. The issue of symmetry naturally arises for the twin data shown in Table 1 in the introduction. Recall that we observe four binary indicators which inform us about each twins' alcohol dependence (A_i) and depression status (D_i). When inspecting Table 1, one notices that counts related by exchanging the index labels 1 and 2 are often very similar.

Let $\mathcal{S}$ be a group of permutations on the index set V. The group $\mathcal{S}$ acts on the set of elementary joint events $\mathcal{I}$ by permuting the components of $i \in \mathcal{I}$. In other words, for $\sigma \in \mathcal{S}$ and $i = (i_v \mid v \in V) \in \mathcal{I}$, we define $\sigma(i) = (i_{\sigma(v)} \mid v \in V)$. This action induces the symmetry model

$$\mathbf{B}(\mathcal{S}) = \{p \in \Delta \mid p_i = p_{\sigma(i)} \ \forall \sigma \in \mathcal{S}\}.$$

Example 16. (*Twin data*). The symmetry group

$$\mathcal{S}_{\text{twin}} = \{(A_1)(A_2)(D_1)(D_2),\ (A_1\,A_2)(D_1\,D_2)\} \tag{24}$$

represents symmetry when exchanging vertex A_1 with A_2, and at the same time exchanging D_1 with D_2. This symmetry corresponds to irrelevance of the labels given to the two twins.

Since the symmetry model is a linear exponential family, the MLE may be computed by simply averaging the empirical cell counts over the orbit induced by the group action:

$$\hat{p}_{\mathcal{S}}(i) = \frac{1}{|S(i)|} \sum_{j \in S(i)} \frac{n(j)}{n}$$

where $S(i) = \{\sigma(i) \mid \sigma \in \mathcal{S}\}$ is the orbit of cell i under the group $\mathcal{S}$, $n(j)$ is the empircal count for cell j, and n is the total sample size. For the twin data the ML estimate is shown in Table 3. The deviance is 4.62 on 6 degrees of freedom, indicating a good fit. We now turn to testing the marginal independence hypothesis mentioned in the introduction, in conjunction with symmetry.

A permutation $\sigma \in \mathcal{S}$ induces a new graph G_σ by renaming vertex $v \in V$ to $\sigma(v) \in V$. In other words, the graph G_σ has the same vertex set V as the original graph $G = (V, E)$ but there is an edge $v \leftrightarrow w$ in G_σ if and only if there is an edge $\sigma^{-1}(v) \leftrightarrow \sigma^{-1}(w)$ in the original graph G. We say that a group of permutations $\mathcal{S}$ leaves the graph G *invariant* if $G_\sigma = G$ for all $\sigma \in \mathcal{S}$, in other words, $\mathcal{S}$ is a subgroup of the automorphism group of G. It follows that no new independences are introduced when imposing symmetry on the distributions in $\mathbf{B}(G)$. We will restrict attention to this case in what follows.

Example 16. (*continued*). Let G be the graph displayed in Figure 2(a), under the variable-vertex correspondence $(1, 2, 3, 4) = (A_1, A_2, D_1, D_2)$ the independence pattern is $A_1 \perp\!\!\!\perp D_2$ and $A_2 \perp\!\!\!\perp D_1$. The group $\mathcal{S}_{\text{twin}}$ given in (24) leaves G invariant.

Theorem 21. *If the symmetry group $\mathcal{S}$ leaves the graph G invariant, then a distribution $p \in \mathbf{B}(G)$ is in the symmetry model $\mathbf{B}(\mathcal{S})$ if and only if the Möbius parameters $q \in Q_G$ for p satisfy that $q_C = q_{\sigma(C)}$ for all connected sets C in G.*

Proof. First, note that under the assumed invariance of the graph, a set $C \subseteq V$ is connected in G if and only if $\sigma(C)$ is connected for all $\sigma \in \mathcal{S}$.

Consider $p \in \mathbf{B}(G) \cap \mathbf{B}(\mathcal{S})$, and let $C \in \mathcal{C}(G)$ and $\sigma \in \mathcal{S}$. Since $i_C = \sigma^{-1}(i)_{\sigma(C)}$ we obtain that

$$q_C = \sum_{i \in \mathcal{I}: i_C = 0} p_i = \sum_{i \in \mathcal{I}: i_C = 0} p_{\sigma^{-1}(i)} = \sum_{j \in \mathcal{I}: j_{\sigma(C)} = 0} p_j = q_{\sigma(C)}.$$

Table 4. ML estimate of the joint distribution under $A_1 \perp\!\!\!\perp D_2$ and $A_2 \perp\!\!\!\perp D_1$ together with the symmetry group $\mathcal{S}_{\text{twin}}$; empirical distribution in parenthesis.

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	0.4612	0.1486	0.1486	0.0957
		(0.4824)	(0.1340)	(0.1541)	(0.0854)
	$A_2 = 1$	0.0249	0.0204	0.0057	0.0104
		(0.0251)	(0.0151)	(0.0117)	(0.0168)
$A_1 = 1$	$A_2 = 0$	0.0249	0.0057	0.0204	0.0104
		(0.0134)	(0.0067)	(0.0134)	(0.0151)
	$A_2 = 1$	0.0100	0.0038	0.0038	0.0054
		(0.0050)	(0.0034)	(0.0067)	(0.0117)

Conversely, assume that the Möbius $q \in Q_G$ satisfy that $q_C = q_{\sigma(C)}$ for all $C \in \mathcal{C}(G)$. Let $D \subseteq V$ be disconnected and uniquely partitioned into inclusion-maximal connected sets as $D = C_1 \dot\cup C_2 \dot\cup \dots \dot\cup C_r$. Then the unique decomposition of $\sigma(D)$ into inclusion-maximal connected sets is given by

$$\sigma(D) = \sigma(C_1) \dot\cup \sigma(C_2) \dot\cup \cdots \dot\cup \sigma(C_r),$$

which implies that

$$q_{\sigma(D)} = q_{\sigma(C_1)} q_{\sigma(C_2)} \cdots q_{\sigma(C_r)} = q_{C_1} q_{C_2} \dots q_{C_r} = q_D.$$

Now consider $i = (0_A, 1_{V \setminus A}) \in \mathcal{I}$. Then $\sigma(i) = (0_{\sigma(A)}, 1_{V \setminus \sigma(A)})$. Using Proposition 6, we obtain that

$$p_{\sigma(i)} = \sum_{B : \sigma(A) \subseteq B} (-1)^{|B \setminus \sigma(A)|} q_B = \sum_{B : A \subseteq \sigma^{-1}(B)} (-1)^{|\sigma^{-1}(B) \setminus A|} q_{\sigma^{-1}(B)} = p_i.$$

□

For a subset $C \subseteq V$, let $S(C) = \{\sigma(C) \mid \sigma \in \mathcal{S}\}$ be the orbit of C.

Corollary 22. *If the symmetry group $\mathcal{S}$ leaves the bi-directed graph G invariant, then the dimension of the marginal independence model with symmetry is*

$$\dim(\mathbf{B}(G) \cap \mathbf{B}(\mathcal{S})) = \sum_{\emptyset \neq C \in \mathcal{C}(G)} 1/|S(C)|.$$

Proof. By dividing through $|S(C)|$, every orbit of connected sets is counted once. □

Corollary 23. *If the symmetry group $\mathcal{S}$ leaves the bi-directed graph G invariant and the marginal independence model with symmetry is restricted to the interior of the probability simplex, then one obtains the curved exponential family $\mathbf{B}_+(G) \cap \mathbf{B}(\mathcal{S})$.*

Proof. The proof is analogous to the proof of Theorem 5.8. □

We define $\hat{n}_{\mathcal{S}} = n \cdot \hat{p}_{\mathcal{S}}$ to be the fitted cell counts under the symmetry model $\mathcal{S}$, which are simply the group averaged cell counts. ML fitting of the model $\mathbf{B}(G) \cap \mathbf{B}(\mathcal{S})$ may be performed by simply applying ICF for fitting $\mathbf{B}(G)$ to $\hat{n}_{\mathcal{S}}$ rather than the observed cell counts. The rationale for this is as follows: let $\mathcal{L}_G(p; \{n(i)\})$ indicate the likelihood for $\mathbf{B}(G)$, evaluated with counts $\{n(i)\}$. If $p \in \mathbf{B}(G) \cap \mathbf{B}(\mathcal{S})$, then $\mathcal{L}_G(p; \{n(i)\}) = \mathcal{L}_G(p; \{\hat{n}_{\mathcal{S}}(i)\})$. Further, $\mathcal{L}_G(p; \{\hat{n}_{\mathcal{S}}(i)\}) = \mathcal{L}_G(\sigma(p); \{\hat{n}_{\mathcal{S}}(i)\})$, for any $\sigma \in \mathcal{S}$, where we define $\sigma(p(i)) = p(\sigma(i))$. Thus the likelihood surface of the independence model $\mathbf{B}(G)$ given the group-averaged counts $\hat{n}_{\mathcal{S}}$ is invariant under permutations $\sigma \in \mathcal{S}$ applied to probability vectors p. It then follows that if p^* is a local maximum of the likelihood function $\mathcal{L}_G(p; \{\hat{n}_{\mathcal{S}}(i)\})$, then so is $\sigma(p^*)$, for any $\sigma \in \mathcal{S}$. Further p^* and $\sigma(p^*)$ are in the same contour of the likelihood function. Consequently if there is at most one local maximum of the likelihood function $\mathcal{L}_G(p; \{\hat{n}_{\mathcal{S}}(i)\})$ in any given contour, then $p^* = \sigma(p^*)$ for all $\sigma \in \mathcal{S}$. Thus a maximum found by ICF when applied to $\hat{n}_{\mathcal{S}}$, is in $\mathbf{B}(\mathcal{S})$, and is thus a maximum of the likelihood for the model of symmetry and independence.

Example 16. (*continued*). Applying ICF to fit the model $A_1 \perp\!\!\!\perp D_2$ and $A_2 \perp\!\!\!\perp D_1$ for the twin data, using the fitted counts from the symmetry model $\mathcal{S}_{\text{twin}}$ resulted in the fitted distribution shown in Table 4. The combined model has a deviance of 16.156 on 2 degrees of freedom, taking the symmetry model given by $\mathcal{S}_{\text{twin}}$ as the alternative. The corresponding p-value of 0.0003 indicates a poor fit and we may safely reject the generating hypothesis represented by the graph in Figure 1(a).

The approach taken here to combining symmetry and independence is analogous to that of Andersson and Madsen (1998) in the Gaussian case. A more general approach would be to apply a symmetry group directly to the Möbius parameters, possibly with the restriction that orbits should only contain parameters corresponding to sets of a given cardinality; this would be more analogous to the work of Højsgaard and Lauritzen (2006).

8. Related Work and Discussion

Several other authors have made use of the Möbius decomposition or similar schemes. Lee (1993) used this decomposition to generate random binary vectors with fixed marginal distributions and specified degrees of association. Ekholm et al. (1995, 2000, 2003) used dependence ratios (see Remark 13) to build association and regression models for multivariate discrete responses. Though Ekholm et al. did not study marginal independence models *per se*, their work on regression models offers one approach to building marginal independence models for mixed continuous and discrete variables, which is an open problem for future work.

Kauermann (1997) developed a parametrization for marginal independence models using the multivariate logistic (m-logit) transformation, which selects the highest order interaction term from every margin. However, the transformation from m-logit parameters to cell probabilities cannot, in general, be computed in closed form. Further, unlike classical log-linear parameters, the valid m-logit parameters may form a complicated subset of $\mathbb{R}^{2^V-1}$ and are not in general variation independent. The m-logit parameterization is a special case of the marginal log-linear framework of Bergsma and Rudas (2002b). In certain cases, such as for Figure 2(a), there may exist a marginal log-linear parameterization for a marginal independence model in which the parameters are variation independent; see Bergsma and Rudas (2002b), Lupparelli and Marchetti (2005). However, there are models for which this approach does not appear to lead to variation independent parametrizations. Specifically, there does not appear to be such a parametrization for the bi-directed chordless five cycle; see Bergsma and Rudas (2002a) for related discussion.

As stated earlier, the problems inherent in expressing marginal independence constraints in terms of a log-linear parametrization over a larger set of variables are part of the general problem of 'lack of upward compatibility': specifically, a log-linear two-way interaction expresses a property of the full joint distribution, and not of the relevant two-way margin. A number of schemes have been proposed for dealing with this problem, in addition to the m-logits mentioned above: see Ip et al. (2003); Streitberg (1999, 1990). These provide alternative parametrizations for the binary bi-directed models introduced here, which may be computed from the fitted distribution, if desired.

Cox (1993) and Cox and Wermuth (1994, 1996) take a different approach to the problem of modelling independence structures similar to Gaussian covariance models. They focus on the quadratic binary exponential distribution, also known as the Boltzmann machine (Hinton and Sejnowski, 1983) or the auto-logistic scheme (Besag, 1974). In this distribution, the absence of a given interaction term does not imply exact marginal independence, but by approximating the marginal distributions via series expansions, it is possible to gauge the size of any such dependence. As Cox notes, the extent to which such marginal approximations are reasonable will depend on the size of the relevant interaction terms.

Acknowledgments. We would like to thank Steen Andersson, Sir David Cox, Steffen Lauritzen, Fero Matuš, Alejandro Murua, Michael Perlman, James Robins, Tamás

Rudas, Milan Studený, Jim Q. Smith, Peter Spirtes, Bernd Sturmfels and Nanny Wermuth for helpful conversations. Kenneth Kendler provided the twin data. Yen-Sheng Chiang and Richard Callahan suggested the analysis of questions relating to Trust. We are particularly grateful to Anders Ekholm for his comments on earlier versions of this paper. This research was supported by the U.S. National Science Foundation (DMS-9972008, DMS-0505612, DMS-0505865), the U.S. National Institutes for Health (R01-HG2362-3), the William and Flora Hewlett Foundation and the Center for Advanced Studies in the Behavioral Sciences at Stanford University where Thomas Richardson was a Fellow from 2003-2004.

Appendix: Likelihood Equations and Hessian calculations

If G is a bi-directed graph with vertex set V, then for an arbitrary subset $A \subseteq V$, let

$$[A]_G = \{C \mid C \text{ is a maximal connected component of } G_A\}.$$

Note that $[A]_G$ forms a partition $A = \bigcup_{C \in [A]_G} C$. For disconnected sets $D \subseteq V$ this partition is the one used in Theorem 8. Since for a connected set $C \subseteq V$ the family $[C]_G$ only comprises one set, namely C itself, we have that under a joint distribution in the model $\mathbf{B}(G)$,

$$q_A = \prod_{C \in [A]_G} q_C, \quad A \subseteq V.$$

Hence for any set A, there is a unique expansion of the joint cell probability p_A^V in terms of the parameters q_C for connected sets C in G,

$$p_A^V = \sum_{B : A \subseteq B} (-1)^{|B \setminus A|} \prod_{C : C \in [B]_G} q_C,$$

recall that $p_A^V = P(X_A = 0, X_{V \setminus A} = 1)$. We call this last expression the *expansion* for p_A^V (under graph G).

Lemma 10. *If C is a connected set in the graph G, then the parameter q_C appears in the expansion for p_A^V if and only if* $\mathrm{Sp}(C) \cap (A \setminus C) = \emptyset$.

Proof. If $\mathrm{Sp}(C) \cap (A \setminus C) = \emptyset$ then $C \cup (A \setminus C)$ forms a disconnected superset of A in which C is a maximal connected component. Hence $C \in [C \cup (A \setminus C)]_G$. If $\mathrm{Sp}(C) \cap (A \setminus C) \neq \emptyset$ then there is a vertex $a \in A \setminus C$ such that $a \in \mathrm{Sp}(C)$. Hence, in any set B containing A and C, there is a maximal connected set $\bar{C} \supseteq C \cup \{a\}$. Hence $C \notin [B]_G$ for any $B \supseteq A$. □

In words, Lemma 10 states that parameter q_C appears in the expansion for p_A^V if and only if every vertex in A that is adjacent to C is already in C. Consequently, $(\partial/\partial q_C) p_A^V = 0$ for any connected set C in G that satisfies $\mathrm{Sp}(C) \cap (A \setminus C) \neq \emptyset$.

Lemma 11. *If* $\mathrm{Sp}(C) \cap (A \setminus C) = \emptyset$ *then*

$$\frac{\partial p_A^V}{\partial q_C} = (-1)^{|C \setminus A|} p_{A \setminus C}^{V \setminus \mathrm{Sp}(C)}.$$

Proof. The claim holds since $\mathrm{Sp}(C) \cap (A \setminus C) = \emptyset$ iff $(A \setminus C) \subseteq (V \setminus \mathrm{Sp}(C))$, and

$$\begin{aligned}
\frac{\partial p_A^V}{\partial q_C} &= \sum_{\substack{B:(A\setminus C)\subseteq B \\ \text{and} B\subseteq V\setminus \mathrm{Sp}(C)}} (-1)^{|(B \dot\cup C)\setminus A|} \prod_{C^*:C^*\in[B]_G} q_{C^*} \\
&= (-1)^{|C\setminus A|} \sum_{\substack{B:(A\setminus C)\subseteq B \\ \text{and} B\subseteq V\setminus \mathrm{Sp}(C)}} (-1)^{|B\setminus A|} \prod_{C^*:C^*\in[B]_G} q_{C^*} \\
&= (-1)^{|C\setminus A|} p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)}.
\end{aligned}$$

□

Corollary 12. *The system of likelihood equations associated with the model* $\mathbf{P}(\mathbf{G})$ *contains an equation*

$$\frac{\partial \ell}{\partial q_C} = \sum_{A:\mathrm{Sp}(C)\cap(A\setminus C)=\emptyset} (-1)^{|C\setminus A|} \frac{n_A^V}{p_A^V} p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)} = 0$$

for every (non-empty) connected set C *in* G.

The likelihood equations can also be expressed in terms of expectations with respect to conditional empirical measures (provided these exist):

$$\mathbb{E}_{X_{V\setminus(\mathrm{Sp}(C)\setminus C)}|X_{\mathrm{Sp}(C)\setminus C}=1}\left[(-1)^{\sum_{i\in C} X_i} P\left(X_{\mathrm{Sp}(C)} \mid X_{V\setminus \mathrm{Sp}(C)}\right)^{-1}\right] = 0$$

where $\mathbb{E}_{X_{V\setminus(\mathrm{Sp}(C)\setminus C)}|X_{\mathrm{Sp}(C)\setminus C}=1}$ is expectation w.r.t. the measure on $\mathcal{I}_{V\setminus(\mathrm{Sp}(C)\setminus C)}$ given by (normalizing) the empirical frequencies in the sub-table in which $X_{\mathrm{Sp}(C)\setminus C} = 1$.

Lemma 13. *Let* C *and* $\bar{C}$ *be connected sets in* G.

(i) *If* $(\mathrm{Sp}(C)\cap(A\setminus C)) \cup \left(\mathrm{Sp}(\bar{C})\cap(A\setminus \bar{C})\right) \neq \emptyset$, *then the second derivative*

$$\frac{\partial}{\partial q_C}\frac{\partial}{\partial q_{\bar{C}}} \log p_A^V = 0.$$

(ii) *If* $(\mathrm{Sp}(C)\cap(A\setminus C)) \cup \left(\mathrm{Sp}(\bar{C})\cap(A\setminus \bar{C})\right) = \emptyset$ *and* $\bar{C}\cap \mathrm{Sp}(C) \neq \emptyset$, *then*

$$\frac{\partial}{\partial q_C}\frac{\partial}{\partial q_{\bar{C}}} \log p_A^V = -(-1)^{|C\setminus A|}(-1)^{|\bar{C}\setminus A|} p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)} p_{A\setminus \bar{C}}^{V\setminus \mathrm{Sp}(\bar{C})} \cdot \frac{1}{(p_A^V)^2}.$$

(iii) *If* $(\mathrm{Sp}(C)\cap(A\setminus C)) \cup \left(\mathrm{Sp}(\bar{C})\cap(A\setminus \bar{C})\right) = \emptyset$ *and* $\bar{C}\cap \mathrm{Sp}(C) = \emptyset$, *then*

$$\begin{aligned}
\frac{\partial}{\partial q_C}\frac{\partial}{\partial q_{\bar{C}}} \log p_A^V = (-1)^{|(C\cup\bar{C})\setminus A)|} p_{A\setminus(C\cup\bar{C})}^{V\setminus \mathrm{Sp}(C\cup\bar{C})} \cdot \frac{1}{p_A^V} \\
- (-1)^{|C\setminus A|}(-1)^{|\bar{C}\setminus A|} p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)} p_{A\setminus \bar{C}}^{V\setminus \mathrm{Sp}(\bar{C})} \cdot \frac{1}{(p_A^V)^2}.
\end{aligned}$$

Proof. This follows from Lemma 11. The first term on the RHS of the equation in (iii) occurs if the derivative of $(\partial/\partial\bar{C})p_{A\setminus C}^{V\setminus \mathrm{Sp}(C)}$ is non-zero, which requires $\bar{C} \subseteq V \setminus \mathrm{Sp}(C)$ and $\mathrm{Sp}(\bar{C}) \cap \big(A \setminus (C \cup \bar{C})\big) = \emptyset$. The second condition is implied by $\mathrm{Sp}(\bar{C}) \cap (A \setminus \bar{C}) = \emptyset$. The first is equivalent to $\bar{C} \cap \mathrm{Sp}(C) = \emptyset$. □

In words, the condition that $\bar{C} \cap \mathrm{Sp}(C) = \emptyset$ requires that C and $\bar{C}$ are disjoint and there is no vertex in C adjacent to a vertex in $\bar{C}$. Note that $\mathrm{Sp}(C) \cap \bar{C} = \emptyset$ if and only if $\mathrm{Sp}(\bar{C}) \cap C = \emptyset$, hence the conditions in (ii) and (iii) are symmetric in C and $\bar{C}$ as required.

The full Hessian may be obtained by summing the expression given in the last Lemma over all sets $A \subseteq V$.

References

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pp. 55–66. New York: Academic Press.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In *Essays in Probability and Statistics*, pp. 1–24. University of North Carolina Press, Chapel Hill, N.C.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist. 1*, 135–141.

Andersson, S. and J. Madsen (1998). Symmetry and lattice conditional independence in a multivariate normal distribution. *Ann. Statist. 26*, 525–572.

Bergsma, W. and T. Rapcsák (2005). An exact penalty method for smooth equality constrained optimization with application to maximum likelihood estimation. Technical Report 1, EURANDOM, Eindhoven. `http://www.eurandom.nl/reports/2005/001WBreport.pdf`.

Bergsma, W. and T. Rudas (2002a). Variation independent parameterizations of multivariate categorical distributions. In C. Cuadras, J. Fortiana, and J. Rodriguez-Lallena (Eds.), *Distributions with given marginals and related topics*, pp. 21–28. Kluwer.

Bergsma, W. P. and T. Rudas (2002b). Marginal models for categorical data. *Ann. Statist. 30*(1), 140–159.

Bertsekas, D. P. (1999). *Nonlinear Programming* (Second ed.). Athena Scientific.

Besag, J. (1974). On spatial-temporal models and Markov fields. In *Transactions of the* 7^{th} *Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 47–55. Academia, Prague.

Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Nat. Acad. Sci. USA 97*, 12182–12186.

Chaudhuri, S., M. Drton, and T. S. Richardson (2007). Estimation of a covariance matrix with zeros. *Biometrika 94*(1), 199–216.

Cox, D. R. (1993). Causality and graphical models. In *Proceedings, 49th Session*, Volume 1 of *Bulletin of the International Statistical Institute*, pp. 363–372.

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

Cox, D. R. and N. Wermuth (1994). A note on the quadratic exponential binary distribution. *Biometrika 81*, 403–408.

Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman and Hall.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist. 8*, 522–539.

Darroch, J. N. and T. P. Speed (1983). Additive and multiplicative models and interactions. *Ann. Statist. 11*(3), 724–738.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B 41*, 1–31.

Diaconis, P. and S. N. Evans (2002). A different construction of Gaussian fields from Markov chains: Dirichlet covariances. *Ann. I. H. Poincaré 38*(6), 863–878.

Drton, M. and M. Eichler (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist. 33*(2), 247–257.

Drton, M. and T. S. Richardson (2003). A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In U. Kjærulff and C. Meek (Eds.), *Proceedings of the* 19th *Conference on Uncertainty in Artificial Intelligence*, pp. 184–191. San Francisco: Morgan Kaufmann.

Edwards, D. M. (2000). *Introduction to Graphical Modelling* (Second ed.). New York: Springer-Verlag.

Ekholm, A., J. Jokinen, J. W. McDonald, and P. W. F. Smith (2003). Joint regression and association modeling of longitudinal ordinal data. *Biometrics 59*(4), 795–803.

Ekholm, A., J. Jokinen, J. W. McDonald, and P. W. F. Smith (2006a). Applying the ejms06-model to the hakim et al. (2003) data. Technical report. http://www.helsinki.fi/ ekholm/hakim.pdf.

Ekholm, A., J. Jokinen, J. W. McDonald, and P. W. F. Smith (2006b). A latent class model for bivariate binary responses from twins. Technical report, University of Southampton, Southampton Statistical Sciences Research Institute (S3RI Methodology Working Papers, M06/10), Southampton, UK. http://eprints.soton.ac.uk/39276/.

Ekholm, A., J. W. McDonald, and P. W. F. Smith (2000). Association models for a multivariate binary response. *Biometrics 56*, 712–718.

Ekholm, A., P. W. F. Smith, and J. W. McDonald (1995). Marginal regression analysis of a multivariate binary response. *Biometrika 82*, 847–854.

Erdös, P. and L. Lovász (1975). Problems and results on 3-chromatic hypergraphs and some related questions. In A. Hajnal, R. Rado, and V. Sós (Eds.), *Infinite and Finite Sets*, pp. 609–628. Amsterdam: North Holland.

Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput. 41*(2), 222–233.

Geiger, D., D. Heckerman, H. King, and C. Meek (2001). Stratified exponential families: graphical models and model selection. *Ann. Statist. 29*(2), 505–529.

Glonek, G. F. V. and P. McCullagh (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B 57*(3), 533–546.

Grzebyk, M., P. Wild, and D. Chouanière (2004). On identification of multi-factor models with correlated residuals. *Biometrika 91*, 141–151.

Haber, M. (1986). Testing for pairwise independence. *Biometrics 42*, 429–435.

Hinton, G. E. and T. J. Sejnowski (1983). Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, New York, pp. 448–453. IEEE.

Højsgaard, S. and S. Lauritzen (2006). Graphical Gaussian models with edge and vertex symmetries. `http://www.stats.ox.ac.uk/∼steffen/papers/rcoxrss.pdf`.

Ip, E., Y. J. Wang, and Y. Yeh (2003). Some equivalence results concerning multiplicative lattice decompositions of multivariate densities. *J. Multivariate Anal. 84*, 403–409.

Jokinen, J. (2006). Fast estimation algorithm for likelihood-based analysis of repeated categorical responses. *Computational Statistics & Data Analysis 51*(3), 1509–1522.

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. *Austral. J. Statist. 39*(3), 261–276.

Kendler, K. S., M. C. Neale, R. C. Kessler, A. C. Kessler, and L. J. Eaves (1992). A population-based twin study of major depression in women. The impact of varying definitions of illness. *Arch. Gen. Psychiatry 49*(4), 257–266.

Kennes, R. and P. Smets (1991). Computational aspects of the Möbius transformation. In P. Bonissone, M. Henrion, L. Kanal, and L. J.F. (Eds.), *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, Amsterdam, pp. 401–416. North Holland.

Knuth, D. E. (1968). *The Art of Computer Programming: Fundamental Algorithms*, Volume 1. Reading, MA: Addison-Wesley.

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

Lauritzen, S. L. (1996). *Graphical Models.* Oxford, UK: Clarendon Press.

Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *Amer. Statist. 47*(3), 209–215.

Levi, M. (1998). A State of Trust. In V. Braithwaite and L. M. (Eds.), *Trust and Governance.* New York: Russell Sage Foundation.

Lupparelli, M. and G. M. Marchetti (2005). Graphical models of marginal independence for categorical variables. In *Convegno SCO 2005*, Padova, pp. 127–132. CLEUP.

Mao, Y., F. R. Kschischang, and B. J. Frey (2004). Convolutional factor graphs as probabilistic models. In U. Kjærulff and C. Meek (Eds.), *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 374–381. San Francisco: Morgan Kaufmann.

Marchetti, G. M. (2006). Independencies induced from a graphical Markov model after marginalization and conditioning: The R package ggm. *Journal of Statistical Software 15*(6).

McCullagh, P. (1989). Models for discrete multivariate responses. In *Proceedings,* 47*th* *Session*, Volume 3 of *Bulletin of the International Statistical Institute*, pp. 407–417.

McCullagh, P. and N. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Moore, A. and M. S. Lee (1998). Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research 8*, 67–91.

Pearl, J. (2000). *Causality.* Cambridge, UK: Cambridge University Press.

Pearl, J. and N. Wermuth (1994). When can association graphs admit a causal interpretation? In *Selecting Models from Data: Artificial Intelligence and Statistics IV*, Volume 89 of *Lecture Notes in Statistics*, pp. 205–214. New York: Springer.

Putnam, R. (2002). *Bowling Alone: The Collapse and Revival of American Community.* Simon and Schuster.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Richardson, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Statist. 30*(1), 145–157.

Richardson, T. S. and P. Spirtes (2002). Ancestral graph Markov models. *Ann. Statist. 30*, 962–1030.

Streitberg, B. (1990). Lancaster interactions revisited. *Ann. Statist. 18*(4), 1878–1885.

Streitberg, B. (1999). Exploring interactions in high-dimensional tables: a bootstrap alternative to log-linear models. *Ann. Statist. 27*(1), 405–413.

Sztompka, P. (2000). *Trust: A Sociological Theory.* Cambridge: Cambridge University Press.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics 32*, 95–108.

Wright, S. (1921). Correlation and causation. *J. Agricultural Research 20*, 557–585.

The University of Chicago, Chicago, U.S.A.

University of Washington, Seattle, U.S.A.
E-mail address: tsr@stat.washington.edu