

# Iterative Conditional Fitting for Discrete Chain Graph Models

Mathias Drton<sup>1</sup>

Department of Statistics, University of Chicago  
5734 S. University Ave, Chicago, IL 60637, U.S.A.  
[drton@uchicago.edu](mailto:drton@uchicago.edu)

**Abstract.** ‘Iterative conditional fitting’ is a recently proposed algorithm that can be used for maximization of the likelihood function in marginal independence models for categorical data. This paper describes a modification of this algorithm, which allows one to compute maximum likelihood estimates in a class of chain graph models for categorical data. The considered discrete chain graph models are defined using conditional independence relations arising in recursive multivariate regressions with correlated errors. This Markov interpretation of the chain graph is consistent with treating the graph as a path diagram and differs from other interpretations known as the LWF and AMP Markov properties.

**Keywords:** Categorical data, chain graph, conditional independence, graphical model

## 1 Introduction

This paper considers models for categorical data that are analogous to Gaussian models induced by systems of linear regression equations with possibly correlated error terms. This analogy is of interest because systems of regression equations appear in many contexts, including structural equation modelling and graphical modelling, see e.g. Koster (1999), Wermuth and Cox (2004). For an example, consider the equations

$$X_1 = \beta_{10} + \varepsilon_1, \tag{1}$$

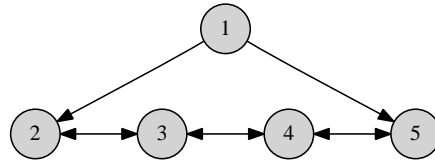
$$X_2 = \beta_{20} + \beta_{21}X_1 + \varepsilon_2, \tag{2}$$

$$X_3 = \beta_{30} + \varepsilon_3, \tag{3}$$

$$X_4 = \beta_{40} + \varepsilon_4, \tag{4}$$

$$X_5 = \beta_{50} + \beta_{51}X_1 + \varepsilon_5, \tag{5}$$

in which the coefficients  $\beta_{ij}$  may be arbitrary real numbers and the error terms  $\varepsilon_i$  have a centered joint multivariate normal distribution with positive definite covariance matrix  $\Omega = (\omega_{ij})$ . This matrix  $\Omega$  is assumed to have a pattern of zero entries such that any pair of error terms is independent except for the pairs  $(\varepsilon_i, \varepsilon_{i+1})$ ,  $i = 2, 3, 4$ , which may have possibly non-zero correlations. These assumptions lead to a particularly structured joint multivariate



**Fig. 1.** Chain graph with directed and bi-directed edges. The use of bi-directed edges is in the tradition of path diagrams employed in structural equation modelling.

normal distribution  $\mathcal{N}(\mu, \Sigma)$  for the random vector  $X = (X_1, \dots, X_5)$ . While the mean vector  $\mu$  is arbitrary, the covariance matrix is of the form

$$\Sigma = \begin{pmatrix} \omega_{11} & \beta_{21}\omega_{11} & 0 & 0 & \beta_{51}\omega_{11} \\ \beta_{21}\omega_{11} & \beta_{21}^2\omega_{11} + \omega_{22} & \omega_{23} & 0 & \beta_{21}\beta_{51}\omega_{11} \\ 0 & \omega_{23} & \omega_{33} & \omega_{34} & 0 \\ 0 & 0 & \omega_{34} & \omega_{44} & \omega_{45} \\ \beta_{51}\omega_{11} & \beta_{21}\beta_{51}\omega_{11} & 0 & \omega_{45} & \beta_{51}^2\omega_{11} + \omega_{55} \end{pmatrix}. \quad (6)$$

The normal model induced by (1)-(5) and the assumptions on the error terms  $\varepsilon_i$  comprises all distributions  $\mathcal{N}(\mu, \Sigma)$  with  $\mu \in \mathbb{R}^5$  arbitrary and  $\Sigma$  of the form (6). This model can be represented using the graph shown in Figure 1. The directed edges in this graph represent covariate-response relations and, in the tradition of the path diagrams employed in the structural equation literature, bi-directed edges represent possible error correlations. The graph in Figure 1 is an instance of a chain graph. Chain graphs may have both oriented and unoriented edges, drawn here as directed and bi-directed edges, subject to acyclicity constraints. A brief introduction to chain graphs is given in §2. A more detailed treatment can be found, for example, in Andersson et al. (2001) and Lauritzen (1996). Note, however, that the so-called LWF and AMP chain graph models discussed in Andersson et al. (2001) and Lauritzen (1996) differ from the models considered in this paper.

It can be shown that the normal model associated with the graph in Figure 1 comprises all multivariate normal distributions in which

$$X_1 \perp\!\!\!\perp (X_3, X_4), \quad X_2 \perp\!\!\!\perp (X_4, X_5) \mid X_1 \quad \text{and} \quad (X_2, X_3) \perp\!\!\!\perp X_5 \mid X_1. \quad (7)$$

Here  $\perp\!\!\!\perp$  denotes marginal or conditional independence depending on whether a conditioning set is specified. Having such a model characterization in terms of the non-parametric concept of conditional independence is useful because it is also meaningful outside the realm of normal distributions. Characterizations such as (7) are available more generally. In particular, if a system of linear regression equations with correlated errors corresponds to a chain graph, then the associated normal model can be characterized in terms of conditional independence. For the details of these results, we refer the reader to Koster (1999) and Richardson and Spirtes (2002).

This paper considers discrete chain graph models for categorical data obtained by imposing conditional independence relations such as (7). We review these models in §3, and in §4 we describe how the recently proposed ‘iterative conditional fitting’ algorithm can be modified for computation of maximum likelihood estimates. Different chain graphs can be Markov equivalent, i.e., lead to the same statistical model. In §5 we discuss how the choice of the graph can affect the computational efficiency of the associated fitting algorithm. Concluding remarks are given in §6.

## 2 Chain graphs

Let  $G = (V, E)$  be a graph with finite vertex set  $V$  and edge set  $E \subseteq (V \times V)$  such that there are no loops, i.e.,  $(v, v) \notin E$  for all  $v \in V$ . Two vertices  $v$  and  $w$  are *adjacent* if  $(v, w) \in E$  or  $(w, v) \in E$ . If  $(v, w) \in E$  and  $(w, v) \in E$ , then the edge  $(v, w) \in E$  is without orientation and, reflective of pictures such as Figure 1, we refer to the edge as *bi-directed*. If  $(v, w) \in E$  but  $(w, v) \notin E$ , then the edge  $(v, w)$  is *directed*. We will also write  $v \rightarrow w$  and  $v \leftrightarrow w$  to indicate directed and bi-directed edges, respectively. If  $v \rightarrow w$  then  $v$  is a *parent* of  $w$ . The set of parents of  $v$  is denoted by  $\text{pa}(v)$ , and for a set of vertices  $\alpha \subseteq V$  we define the parents as

$$\text{pa}(\alpha) = \{w \in V \mid \exists v \in \alpha : w \rightarrow v \text{ in } G\}.$$

A sequence of distinct vertices  $\langle v_0, \dots, v_k \rangle$  is a *path* if  $v_{i-1}$  and  $v_i$  are adjacent for all  $1 \leq i \leq k$ . A path  $\langle v_0, \dots, v_k \rangle$  is a *semi-directed cycle* if  $(v_{i-1}, v_i) \in E$  for all  $0 \leq i \leq k$  and at least one of the edges is directed as  $v_{i-1} \rightarrow v_i$ . Here,  $v_{-1} \equiv v_k$ . If the graph  $G$  has no semi-directed cycles, then  $G$  is a *chain graph*.

Define two vertices  $v_0$  and  $v_k$  in a chain graph  $G$  to be equivalent if there exists a bi-directed path from  $v_0$  to  $v_k$ , i.e., a path  $\langle v_0, \dots, v_k \rangle$  such that  $v_i \leftrightarrow v_{i+1}$  for all  $0 \leq i \leq k-1$ . The equivalence classes under this equivalence relation are the *chain components* of  $G$ . For example, the chain graph in Figure 1 has the chain components  $\{1\}$  and  $\{2, 3, 4, 5\}$ . The chain components  $(\tau \mid \tau \in \mathcal{T})$  yield a partitioning of the vertex set

$$V = \bigcup_{\tau \in \mathcal{T}} \tau,$$

and the subgraph  $G_\tau$  induced by each chain component  $\tau$  is a connected graph with exclusively bi-directed edges. Moreover, the directed edges between two chain components  $\tau_1$  and  $\tau_2$  all have the same direction, i.e., if  $(v_1, v_2) \in \tau_1 \times \tau_2$  and  $(w_1, w_2) \in \tau_1 \times \tau_2$  are two pairs of adjacent vertices, then either  $v_1 \rightarrow v_2$  and  $w_1 \rightarrow w_2$ , or  $v_2 \rightarrow v_1$  and  $w_2 \rightarrow w_1$ .

### 3 Discrete chain graph models of multivariate regression type

Let  $X = (X_v | v \in V)$  be a discrete random vector whose elements correspond to the vertices of a chain graph  $G = (V, E)$ . The graph  $G$  determines a list of conditional independence statements such as (7). The details of the process of determining the conditional independence statements are reviewed, for example, in Drton (2008). Let component  $X_v$  take values in  $[d_v] = \{1, \dots, d_v\}$ , and define  $\mathcal{I} = \times_{v \in V} [d_v]$ . For  $i = (i_v | v \in V) \in \mathcal{I}$ , let

$$p(i) = P(X = i) = P(X_v = i_v \text{ for all } v \in V). \quad (8)$$

The joint distribution of  $X$  is determined by the array of probabilities  $p = (p(i) | i \in \mathcal{I})$ , which is in the  $(\prod_{v \in V} d_v) - 1$  dimensional probability simplex  $\Delta \subset \mathbb{R}^{\mathcal{I}}$ . Hence, the discrete chain graph model associated with  $G$  corresponds to a subset  $\mathcal{P}(G) \subset \Delta$ , which comprises exactly those arrays of probabilities  $p$  that lead to the desired conditional independence relations for the random vector  $X$ .

An array  $p \in \mathcal{P}(G)$  obeys a factorization over the chain components of  $G$ . For a chain component  $\tau \subseteq V$ , let

$$p(i_\tau | i_{\pi(\tau)}) = P(X_\tau = i_\tau | X_{\pi(\tau)} = i_{\pi(\tau)}), \quad (9)$$

where  $\pi(\tau)$  is the union of all chain components  $\tau'$  in  $G$  that contain a vertex  $w$  that is a parent of a vertex in  $\tau$ , i.e., all  $\tau'$  such that  $\tau' \cap \text{pa}(\tau) \neq \emptyset$ . If  $\pi(\tau) = \emptyset$ , then (9) refers to an unconditional probability. It then holds that each  $p \in \mathcal{P}(G)$  factors as

$$p(i) = \prod_{\tau \in \mathcal{T}} p(i_\tau | i_{\pi(\tau)}), \quad i \in \mathcal{I}. \quad (10)$$

The factorization in (10) is of the type usually encountered in directed graphical models (also known as Bayesian networks) but operates on the level of chain components rather than individual vertices. The factorization for the chain graph in Figure 1 is of the form

$$p(i) = p(i_1)p(i_2, i_3, i_4, i_5 | i_1), \quad i \in \mathcal{I}. \quad (11)$$

The conditional independence relations that need to hold in an array of probabilities  $p$  in order for it to be in the model  $\mathcal{P}(G)$  lead to constraints on the conditional probabilities  $p(i_\tau | i_{\pi(\tau)})$ . Drton (2008) describes a change of conditional probability coordinates that simplifies these constraints and yields in particular that the positive distributions in the model  $\mathcal{P}(G)$  form a curved exponential family. This ensures regular large-sample asymptotics such as asymptotically normal maximum likelihood estimators.

Consider a chain component  $\tau \in \mathcal{T}$  and let  $i_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$ . For a non-empty subset  $\alpha \subseteq \tau$ , define  $\mathcal{J}_\alpha = \times_{v \in \alpha} [d_v - 1]$ . The set  $[d_v - 1] = \{1, \dots, d_v - 1\}$

is the range for random variable  $X_v$  but with the highest-numbered element  $d_v$  removed. (Any other element could be chosen as baseline and be removed instead.) For each  $j_\alpha \in \mathcal{J}_\alpha$ , let

$$q(j_\alpha | i_{\pi(\tau)}) = P(X_\alpha = j_\alpha | X_{\pi(\tau)} = i_{\pi(\tau)}).$$

The probabilities  $q(j_\alpha | i_{\pi(\tau)})$ ,  $\emptyset \neq \alpha \subseteq \tau$ ,  $j_\alpha \in \mathcal{J}_\alpha$ , can be shown to be in one-to-one correspondence to the probabilities  $p(i_\tau | i_{\pi(\tau)})$ ,  $i_\tau \in \mathcal{I}_\tau$ . This gives the above mentioned change of coordinates that simplifies the considered conditional independence relations to the form in Theorem 1.

A subset  $\alpha \subseteq \tau$  is *disconnected* if there are two distinct vertices  $v, w \in \alpha$  such that no path from  $v$  to  $w$  in  $G$  has all its vertices in  $\alpha$ . Otherwise,  $\alpha$  is a *connected* set. A disconnected set  $\delta \subseteq \tau$  can be partitioned uniquely into inclusion-maximal disjoint connected sets  $\gamma_1, \dots, \gamma_r \subseteq \tau$ ,

$$\delta = \gamma_1 \cup \gamma_2 \cup \dots \cup \gamma_r. \quad (12)$$

**Theorem 1 (Drton, 2008).** *Let  $G$  be a chain graph with chain components  $(\tau | \tau \in \mathcal{T})$ . An array  $p$  in the probability simplex  $\Delta$  belongs to the discrete chain graph model  $\mathcal{P}(G)$  if and only if the following three conditions hold:*

- (i) *The components of  $p$  factor as in (10).*
- (ii) *For all  $\tau \in \mathcal{T}$  and  $i_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$ , it holds that*

$$q(j_\delta | i_{\pi(\tau)}) = q(j_{\gamma_1} | i_{\pi(\tau)})q(j_{\gamma_2} | i_{\pi(\tau)}) \cdots q(j_{\gamma_r} | i_{\pi(\tau)})$$

*for every disconnected set  $\delta \subseteq \tau$  and  $j_\delta \in \mathcal{J}_\delta$ . Here  $\gamma_1, \dots, \gamma_r \subseteq \tau$  are the inclusion-maximal connected sets in (12).*

- (iii) *For all  $\tau \in \mathcal{T}$ , connected subsets  $\gamma \subseteq \tau$  and  $j_\gamma \in \mathcal{J}_\gamma$ , it holds that*

$$q(j_\gamma | i_{\pi(\tau)}) = q(j_\gamma | k_{\pi(\tau)})$$

*for every pair  $i_{\pi(\tau)}, k_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$  such that  $i_{\text{pa}(\gamma)} = k_{\text{pa}(\gamma)}$ .*

*Example 1.* For the graph from Figure 1, Theorem 1 only constrains the conditional probabilities  $p(i_2, i_3, i_4, i_5 | i_1)$  for the second chain component  $\{2, 3, 4, 5\}$ . The constraints from condition (ii) are

$$q(j_2, j_4 | i_1) = q(j_2 | i_1)q(j_4 | i_1), \quad (13)$$

$$q(j_2, j_5 | i_1) = q(j_2 | i_1)q(j_5 | i_1), \quad (14)$$

$$q(j_3, j_5 | i_1) = q(j_3 | i_1)q(j_5 | i_1), \quad (15)$$

$$q(j_2, j_3, j_5 | i_1) = q(j_2, j_3 | i_1)q(j_5 | i_1), \quad \text{and} \quad (16)$$

$$q(j_2, j_4, j_5 | i_1) = q(j_2 | i_1)q(j_4, j_5 | i_1), \quad (17)$$

for all  $i_1 \in [d_1]$  and  $j_{2345} \in \mathcal{J}_{2345}$ . Condition (iii) leads to the constraints

$$q(j_3 | i_1) = q(j_3 | k_1), \quad (18)$$

$$q(j_4 | i_1) = q(j_4 | k_1), \quad \text{and} \quad (19)$$

$$q(j_3, j_4 | i_1) = q(j_3, j_4 | k_1), \quad (20)$$

for all  $i_1 < k_1 \in [d_1]$  and  $j_{34} \in \mathcal{J}_{34}$ .

## 4 Iterative conditional fitting

Suppose  $X^{(1)}, \dots, X^{(n)}$  are a sample of independent and identically distributed random vectors taking values in  $\mathcal{I} = \times_{v \in V} [d_v]$ . Suppose further that the probability array  $p$  for the joint distribution of the random vectors  $X^{(k)}$  is in a chain graph model  $\mathcal{P}(G)$ . If we define the counts

$$n(i) = \sum_{k=1}^n 1_{\{X^{(k)}=i\}}, \quad i \in \mathcal{I},$$

then the *likelihood function* of  $\mathcal{P}(G)$  is equal to

$$L(p) = \prod_{i \in \mathcal{I}} p(i)^{n(i)}.$$

For  $\alpha \subseteq V$  and  $i_\alpha \in \mathcal{I}_\alpha$ , define the marginal counts

$$n(i_\alpha) = \sum_{j \in \mathcal{I}: j_\alpha = i_\alpha} n(j),$$

and the marginal probabilities

$$p(i_\alpha) = P(X_\alpha = i_\alpha) = \sum_{j \in \mathcal{I}: j_\alpha = i_\alpha} p(j).$$

Using the factorization in (10), the *log-likelihood function*  $\ell(p) = \log L(p)$  can be written as the sum  $\ell(p) = \sum_{\tau \in \mathcal{T}} \ell_\tau(p)$ , where

$$\ell_\tau(p) = \sum_{i_{\pi(\tau)}} \sum_{i_\tau} n(i_\tau, i_{\pi(\tau)}) \log p(i_\tau | i_{\pi(\tau)}) \quad (21)$$

are the *component log-likelihood functions*. For different chain components, the conditional probabilities  $p(i_\tau | i_{\pi(\tau)})$  are variation-independent. We can thus maximize  $\ell$  over  $\mathcal{P}(G)$  by separately maximizing the component log-likelihood functions over their respective domains.

*Example 2.* For the graph from Figure 1 we obtain that  $\ell(p) = \ell_1(p) + \ell_{2345}(p)$  with component log-likelihood functions

$$\ell_1(p) = \sum_{i_1 \in [d_1]} n(i_1) \log p(i_1), \quad (22)$$

$$\ell_{2345}(p) = \sum_{i \in \mathcal{I}} n(i) \log p(i_2, i_3, i_4, i_5 | i_1). \quad (23)$$

The function  $\ell_1$  is maximized by  $\hat{p}(i_1) = n(i_1)/n$  and only the maximization of  $\ell_{2345}$  presents a challenge.

Since the component log-likelihood function  $\ell_\tau$  in (21) is a sum over  $i_{\pi(\tau)}$ , each term of this sum can be maximized separately if one has variation-independence of the probabilities appearing in the different terms.

**Proposition 1.** *Suppose  $\tau$  is a chain component of the chain graph  $G$  such that  $\text{pa}(v) = \pi(\tau)$  for all  $v \in \tau$ . If  $i_{\pi(\tau)}, j_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$  and  $i_{\pi(\tau)} \neq j_{\pi(\tau)}$ , then the two arrays of conditional probabilities  $(p(i_\tau | i_{\pi(\tau)}) | i_\tau \in \mathcal{I}_\tau)$  and  $(p(i_\tau | j_{\pi(\tau)}) | i_\tau \in \mathcal{I}_\tau)$  are variation-independent.*

*Proof.* Since  $\text{pa}(v) = \pi(\tau)$  for all  $v \in \tau$ , condition (iii) in Theorem 1 is void. Condition (ii) constrains each one of the arrays  $(p(i_\tau | i_{\pi(\tau)}) | i_\tau \in \mathcal{I}_\tau)$  and  $(p(i_\tau | j_{\pi(\tau)}) | i_\tau \in \mathcal{I}_\tau)$  separately.  $\square$

Clearly, Proposition 1 applies only in very special cases. It does not apply, for instance, to the chain component  $\{2, 3, 4, 5\}$  of the graph from Figure 1 because  $\text{pa}(2) = \{1\}$  differs from  $\text{pa}(3) = \emptyset$ .

Different approaches can be taken for maximization of a component log-likelihood function  $\ell_\tau$  in an arbitrarily structured chain graph. One approach is to express  $\ell_\tau$  in terms of parametrizations and then apply routines for unconstrained numerical optimization. Note, however, that care must be taken to avoid issues due to the fact that the involved parameters are generally variation-dependent. Here we will take an alternative approach by generalizing the ‘iterative conditional fitting’ (ICF) algorithm described in Drton and Richardson (2008a). This algorithm was proposed for binary marginal independence models and has a Gaussian version discussed in Chaudhuri et al. (2007). The marginal independence models studied in Drton and Richardson (2008a) are special cases of the chain graph models considered here. They are obtained from chain graphs with only one chain component, in which case conditions (i) and (iii) in Theorem 1 are void.

Generalized ICF for maximization of  $\ell_\tau$  from (21) starts with a choice of feasible estimates of the probabilities  $p(i_\tau | i_{\pi(\tau)})$ . These estimates are then improved iteratively. Each iteration cycles through all vertices in  $\tau$  and when considering vertex  $v \in \tau$  an update step with three parts is performed:

- (a) Use the current feasible estimates to compute the conditional probabilities

$$p(i_{\tau \setminus \{v\}} | i_{\pi(\tau)}),$$

in which the variable  $X_v$  is marginalized out.

- (b) Holding the probabilities computed in (a) fixed, solve a convex optimization problem to find updated estimates of the conditional probabilities

$$p(i_v | i_{(\tau \setminus \{v\}) \cup \pi(\tau)}).$$

- (c) Compute new feasible estimates according to

$$p(i_{\tau \setminus \{v\}} | i_{\pi(\tau)}) = p(i_v | i_{(\tau \setminus \{v\}) \cup \pi(\tau)}) p(i_{\tau \setminus \{v\}} | i_{\pi(\tau)}).$$

The update step (a)-(c) mirrors the corresponding step in the original ICF algorithm, however, we now condition on the variables  $X_{\pi(\tau)}$  throughout.

It remains to explain which convex optimization problem has to be solved in part (b). The problem is to maximize  $\ell_\tau$ , treating the conditional probabilities from part (a) as fixed quantities and imposing the constraints from Theorem 1(ii) and (iii). To see the convexity of the problem, note that for fixed values of  $p(i_\tau \setminus \{v\} | i_{\pi(\tau)})$  the function  $\ell_\tau$  is a concave function of the probabilities  $p(i_v | i_{(\tau \setminus \{v\}) \cup \pi(\tau)})$ . Moreover, for fixed  $p(i_\tau \setminus \{v\} | i_{\pi(\tau)})$ , the constraints in Theorem 1(ii) and (iii) are linear in  $p(i_v | i_{(\tau \setminus \{v\}) \cup \pi(\tau)})$ . Thus the feasible set for the problem is convex.

*Example 3.* We illustrate the outlined algorithm for the chain graph  $G$  in Figure 1 and the component log-likelihood function  $\ell_{2345}$  from (23). For simplicity, we assume all five variables to be binary, i.e.,  $d_v = 2$  for  $v = 1, \dots, 5$ . Up to symmetry, there are only two different update steps in ICF, namely, the one for  $v = 2$  and the one for  $v = 3$ .

*Update step for  $v = 2$ :* In part (a) we compute the 16 conditional probabilities

$$p(i_3, i_4, i_5 | i_1), \quad i_1, i_3, i_4, i_5 = 1, 2.$$

These are then treated as fixed in part (b), in which we maximize

$$\sum_{i_1, \dots, i_5=1}^2 n(i_1, i_2, i_3, i_4, i_5) \log p(i_2 | i_1, i_3, i_4, i_5) \quad (24)$$

with respect to  $p(i_2 | i_1, i_3, i_4, i_5)$ . This maximization is done under constraints derived from (13), (14), (16) and (17). Since the probabilities  $p(i_3, i_4, i_5 | i_1)$  are fixed, (15) is preserved automatically when updating the probabilities  $p(i_2 | i_1, i_3, i_4, i_5)$ . Moreover, since  $\text{pa}(2) = \pi(\tau)$  for  $\tau = \{2, 3, 4, 5\}$ , conditions (18)-(20) do not constrain the probabilities  $p(i_2 | i_1, i_3, i_4, i_5)$ . The important point is that (13), (14), (16) and (17) are linear constraints in  $p(i_2 | i_1, i_3, i_4, i_5)$ . The second factor on the right hand side of these equations is a function of the fixed quantities  $p(i_3, i_4, i_5 | i_1)$  and thus also fixed. All other terms are linear combinations of  $p(i_2 | i_1, i_3, i_4, i_5)$ . For instance, (17) is linearized by writing

$$q(j_2, j_4, j_5 | i_1) = \sum_{i_3=1}^2 p(j_2 | i_1, i_3, j_4, j_5) p(i_3, j_4, j_5 | i_1), \quad (25)$$

$$q(j_2 | i_1) = \sum_{i_3, i_4, i_5=1}^2 p(j_2 | i_1, i_3, i_4, i_5) p(i_3, i_4, i_5 | i_1), \quad (26)$$

and noting that the probabilities  $p(i_3, i_4, i_5 | i_1)$  are fixed quantities. In the resulting constrained maximization of (24) the two terms corresponding

to  $i_1 = 1, 2$  can in fact be maximized separately because none of the constraints obtained from (13), (14), (16) and (17) involve two conditional probabilities  $p(i_2 | i_1, i_3, i_4, i_5)$  and  $p(k_2 | k_1, k_3, k_4, k_5)$  with  $i_1 \neq k_1$ .

*Update step for  $v = 3$ :* In part (a) we compute again 16 conditional probabilities, namely,

$$p(i_2, i_4, i_5 | i_1), \quad i_1, i_2, i_4, i_5 = 1, 2.$$

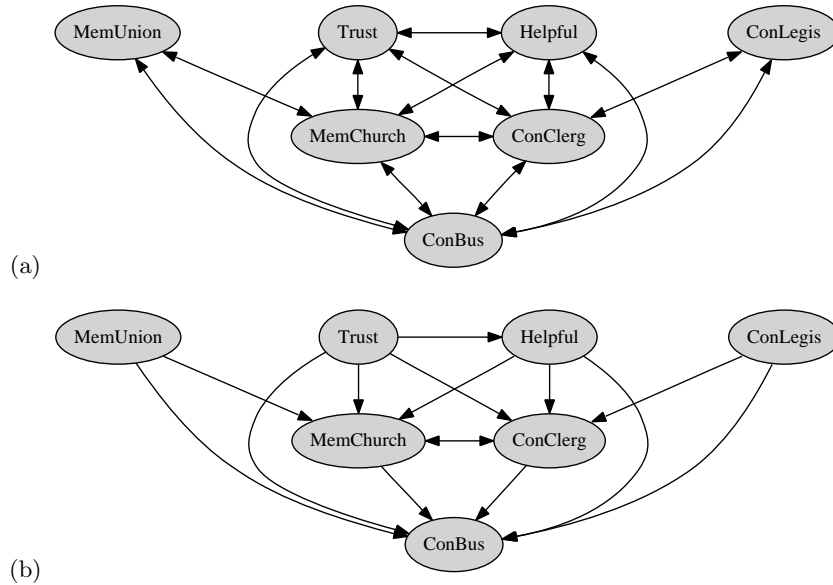
The objective function for part (b) is analogous to (24) but the probabilities  $p(i_2 | i_1, i_3, i_4, i_5)$  are replaced by  $p(i_3 | i_1, i_2, i_4, i_5)$ . The relevant constraints are now (15), (16), (18) and (20). These can be linearized as in (25) and (26). The equations derived from (18) and (20) now involve  $p(i_2 | i_1, i_3, i_4, i_5)$  and  $p(k_2 | k_1, k_3, k_4, k_5)$  with  $i_1 < k_1$ . Hence, the optimization problem cannot be decomposed any further by splitting the strata  $i_1 = 1$  and  $i_1 = 2$ .

In the update step for  $v = 2$  in the above example, the terms in (24) indexed by different values of  $i_1$  can be maximized separately. This observation is general in that analogous decompositions are possible if a vertex  $v \in \tau$  satisfies  $\text{pa}(v) = \pi(\tau)$ . This follows because  $\text{pa}(v) = \pi(\tau)$  implies that condition (iii) in Theorem 1 remains void; also recall Proposition 1.

## 5 Fitting marginal independence models via chain graphs

Graphical models based on graphs with directed edges generally lead to the problem of Markov equivalence, which arises if two different graphs induce the same statistical model. While such Markov equivalence poses challenges for the statistical interpretation of graphs, it can sometimes be exploited for more efficient computation of maximum likelihood estimates. Fitting algorithms, such as ICF, are often specified in terms of the underlying graph. The idea is then to find, among several Markov equivalent graphs, the one for which the associated fitting algorithm runs the fastest. Drton and Richardson (2008b) pursued this idea in the case of marginal independence models for the multivariate normal distribution and presented an algorithm that converts a graph with bi-directed edges into a Markov equivalent graph more suitable for optimization of the likelihood function. As we illustrate next, these constructions are also useful in the discrete case.

Table 2 in Drton and Richardson (2008a) presents data on seven binary variables from the US General Social Survey (sample size  $n = 13\,486$ ). The data and software for their analysis are available for download at a supporting website for that article. A backward selection among bi-directed graphs for these data yields the graph  $G_a$  in Figure 2(a), which is identical to the one in Figure 4(a) in Drton and Richardson (2008a). Since all edges are bi-directed,



**Fig. 2.** (a) Bi-directed graph  $G_a$  for data from the US General Social Survey. (b) Chain graph  $G_b$  that is Markov equivalent to the bi-directed graph  $G_a$ .

the model  $\mathcal{P}(G_a)$  can be characterized by marginal independences. The ICF algorithm for  $G_a$  iteratively estimates conditional probabilities of each of the seven variables given the remaining six variables. Running the algorithm for the given data, the deviance of  $\mathcal{P}(G_a)$  was found to be 32.67 over 26 degrees of freedom, when compared with the saturated model of no independence. The asymptotic chi-square p-value is 0.172.

Using the results in Drton and Richardson (2008b), it is easily seen that the bi-directed graph  $G_a$  in Figure 2(a) is Markov equivalent to the chain graph  $G_b$  in Figure 2(b), i.e.,  $\mathcal{P}(G_a) = \mathcal{P}(G_b)$ . When passing from  $G_a$  to  $G_b$  all but one of the bi-directed edges were substituted by directed edges. The remaining bi-directed edge in  $G_b$  between *MemChurch* and *ConClerg* cannot be replaced by a directed edge without destroying Markov equivalence to  $G_a$ . The graph  $G_b$  has six chain components, namely,

$$\begin{aligned} \tau_1 &= \{Trust\}, & \tau_2 &= \{Helpful\}, & \tau_3 &= \{MemUnion\}, \\ \tau_4 &= \{ConLegis\}, & \tau_5 &= \{MemChurch, ConClerg\}, & \tau_6 &= \{ConBus\}. \end{aligned}$$

The factorization (10) takes the form

$$\begin{aligned} p(i) &= p(i_T)p(i_H | i_T)p(i_{MU})p(i_{CL})p(i_{MC}, i_{CC} | i_T, i_H, i_{MU}, i_{CL}) \times \\ & \quad p(i_{CB} | i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL}), \end{aligned} \quad (27)$$

where the indices identify the variables via the capital letters in their names. The factorization in (27) reveals several closed-form maximum likelihood es-

timates, namely,

$$\begin{aligned}\hat{p}(i_T) &= \frac{n(i_T)}{n}, \\ \hat{p}(i_H | i_T) &= \frac{n(i_T, i_H)}{n(i_T)}, \\ \hat{p}(i_{MU}) &= \frac{n(i_{MU})}{n}, \\ \hat{p}(i_{CL}) &= \frac{n(i_{CL})}{n},\end{aligned}$$

and

$$\hat{p}(i_{CB} | i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL}) = \frac{n(i_{CB}, i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL})}{n(i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL})};$$

see Drton (2008) where this observation is made in generality. The problem of computing maximum likelihood estimates in  $\mathcal{P}(G_b)$  can thus be reduced to the simpler problem of maximizing the component log-likelihood function  $\ell_{\tau_5}$ . Moreover, since  $\tau_5$  is a complete set (there is an edge between all of its elements),  $\ell_{\tau_5}$  and thus also the likelihood function of  $\mathcal{P}(G_b)$  have a unique local and global maximum if all counts are positive as is the case for the considered data; see again Drton (2008) for a general version of this result.

The maximization of  $\ell_{\tau_5}$  can be effected using the generalization of ICF developed in §4. It requires alternating between two update steps that estimate the conditional probabilities of *MemChurch* and of *ConClerg* given the respective five remaining variables in the components  $\tau_m$  with  $m \leq 5$ . The variable *ConBus* in  $\tau_6$  is marginalized out in this computation. The constraints in part (b) of the update step for *MemChurch* arise from condition (iii) in Theorem 1. They take the form

$$\begin{aligned}\sum_{i_{CC}=1}^2 p(j_{MC} | i_{CC}, i_T, i_H, i_{MU}, i_{CL})p(i_{CC} | i_T, i_H, i_{MU}, i_{CL}) = \\ \sum_{i_{CC}=1}^2 p(j_{MC} | i_{CC}, i_T, i_H, i_{MU}, k_{CL})p(i_{CC} | i_T, i_H, i_{MU}, k_{CL}),\end{aligned}\quad (28)$$

for  $j_{MU} = 1$ ,  $i_{CL} = 1$ ,  $k_{CL} = 2$  and arbitrary combinations of  $i_T$ ,  $i_H$  and  $i_{MU}$ . The variables being binary, (28) corresponds to a total of eight constraints. The symmetry present in  $G_b$  implies that the update step for *ConClerg* is analogous to that for *MemChurch*.

Running the ICF algorithm associated with  $G_b$  as opposed to  $G_a$  produced the same results with substantial reduction in computation time. When implementing the two algorithms in the statistical programming environment ‘R’ and using the same routine to solve the constrained optimization problem arising in the parts (b) of the respective update steps, we found that using  $G_b$  reduced the running time by a factor of about 70.

## 6 Conclusion

We have described a simple modification of the ‘iterative conditional fitting’ (ICF) algorithm proposed for marginal independence models in Drton and Richardson (2008a). This modification allows one to fit models associated with chain graphs. In future work it would be interesting to compare or even combine the ICF approach with other approaches to computation of maximum likelihood estimates such as that of Lupparelli et al. (2008).

As illustrated in §5, fitting algorithms for graphical models may take different forms for Markov equivalent graphs, and choosing the ‘right’ graph from a Markov equivalence class can be crucial for computationally efficient model fitting. Graphical constructions relevant for the models considered in this paper are given in Drton and Richardson (2008b) and Ali et al. (2005). However, these constructions generally do not return a chain graph but rather a graph from the larger class of ancestral graphs introduced in Richardson and Spirtes (2002). Generalizing ICF and other fitting algorithms to cover discrete ancestral graph models is an interesting topic for future research.

## References

- ALI, R. A., RICHARDSON, T. S., SPIRTEs, P. and ZHANG, J. (2005): Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In: F. Bacchus and T. Jaakkola (Eds.): *Proc. 21st Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Corvallis, Oregon, 10-17.
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001): Alternative Markov properties for chain graphs. *Scand. J. Statist.* 28, 33-85.
- CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007): Estimation of a covariance matrix with zeros. *Biometrika* 94, 199-216.
- DRTON, M. (2008): Discrete chain graph models. Manuscript.
- DRTON, M. and RICHARDSON, T. S. (2008a): Binary models for marginal independence. *J. R. Stat. Soc. Ser. B.* 70, 287-309. (Software and data available at <http://www.blackwellpublishing.com/rss/Volumes/Bv70p2.htm>)
- DRTON, M. and RICHARDSON, T. S. (2008b): Graphical methods for efficient likelihood inference in Gaussian covariance models. Preprint available at [arXiv:0708.1321](https://arxiv.org/abs/0708.1321).
- KOSTER, J. T. A. (1999): On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems with correlated errors. *Scand. J. Statist.* 26, 413-431.
- LAURITZEN, S. L. (1996): *Graphical Models*. Oxford University Press, Oxford, 1996.
- LUPPARELLI, M., MARCHETTI, G. M. and BERGSMA, W. P. (2008): Parameterizations and fitting of bi-directed graph models to categorical data. Preprint available at [arXiv:0801.1440](https://arxiv.org/abs/0801.1440).
- RICHARDSON, T. S. and SPIRTEs, P. (2002): Ancestral graph Markov models. *Ann. Statist.* 30, 962-1030.
- WERMUTH, N. and COX, D. R. (2004): Joint response graphs and separation induced by triangular systems. *J. R. Stat. Soc. Ser. B* 66, 687-717.