Marginal conditional independence models with application to graphical modeling

Tamás Rudas
Department of Statistics, Faculty of Social Sciences
Eötvös Loránd University, Budapest
rudas@tarki.hu

Wicher Bergsma
Department of Statistics
London School of Economics and Political Science
W.P.Bergsma@lse.ac.uk

Renáta Németh
Department of Statistics, Faculty of Social Sciences
Eötvös Loránd University, Budapest
nmthrnt@freemail.hu

June 6, 2009

Abstract

Conditional independence models are defined by a set of conditional independence restrictions and play an important role in many statistical applications, especially, but not only, graphical modeling. In this paper we identify a subclass of these models which are hierarchical marginal log-linear, as defined by Bergsma and Rudas (2002a). Such models are smooth, which implies the applicability of standard asymptotic theory and simplifies interpretation. Furthermore, we give a marginal log-linear parameterization and a minimal specification of the models in the subclass, which implies the applicability of standard methods to compute maximum likelihood estimates and simplifies the calculation of the degrees of freedom for the model. We illustrate the utility of our results by applying them to certain block recursive Markov models associated with chain graphs.

1 Introduction

Conditional independence models have received considerable attention in the literature of the past few decades, see for example Studeny (2004), Drton, Sturmfels, and Sullivant (2009) and references therein. A conditional independence model consists of one or more conditional independence restrictions on a set of random variables. Graphical models (see, e.g., Cox and Wermuth (1996), Lauritzen (1996) and the references in Section 3) are perhaps the most important statistical models defined using conditional independencies.

Conditional independence models may show unexpected behavior. To illustrate, consider the following two examples for random variables A, B, and C. Firstly, the intersection of $A \perp\!\!\!\perp C$ and $A \perp\!\!\!\perp B \,|\, C$ can be verified to be equivalent to $A \perp\!\!\!\perp BC$. Secondly, if C is dichotomous, the intersection of $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \,|\, C$ is equivalent to the union of $A \perp\!\!\!\perp BC$ and $B \perp\!\!\!\perp AC$, that has non-standard behavior (see the discussion in Example 4).

This paper considers strictly positive distributions on contingency tables and identifies a subclass of conditional independence models which belong to the class of marginal log-linear models developed by Bergsma and Rudas (2002a) – hereinafter referred to as BR. Such models are smooth, a characteristics that aids the interpretation of the models, and guarantees the applicability of standard asymptotic theory. The aforementioned example concerning the intersection of $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \,|\, C$ is nonsmooth at any distribution satisfying mutual independence of the three variables. Our main result, given in Section 2, is a combinatorial condition on the sets of variables involved in the conditional independence restrictions, that guarantees that the model is a hierarchical marginal log-linear model and hence smooth. Furthermore, a minimal specification of such models is obtained, as well as a marginal log-linear parameterization. The minimal specification is necessary to apply the fitting algorithms described by Lang and Agresti (1994), Bergsma (1997) and Bergsma and Rapcsak (2006), and allows for easy computation of the degrees of freedom for the model. The second algorithm is used in Bergsma, Croon, and Hagenaars (2009). A parameterization is also necessary to use the algorithm described by Glonek and McCullagh (1995). The proofs of these results uses theorems of BR but do not immediately follow from them.

In Section 3, we use the main result to prove that block recursive Markov models associated with chain graphs, called Type IV models by Drton (2009) (see also Andersson, Madigan, & Perlman, 2001) are smooth and give their hierarchical marginal log-linear parameterizations. We also give such pa-

rameterizations for the so-called LWF block recursive models (see Lauritzen & Wermuth, 1989 and Frydenberg, 1990).

2 Conditional independence models as marginal log-linear models

Let $\mathcal{V}$ be a set of categorical variables and let $\mathcal{P}$ denote the set of strictly positive joint probability distributions for $\mathcal{V}$. For pairwise disjoint subsets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ of $\mathcal{V}$, with $\mathcal{A}$ and $\mathcal{B}$ nonempty, a conditional independence restriction is

$$\mathcal{Q}_1 = \{P \in \mathcal{P} : \mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C}(P)\}\,. \tag{1}$$

A conditional independence model is defined as an intersection of such restrictions. If for $i = 1, \ldots, k$, if $\mathcal{A}_i \neq \emptyset$, $\mathcal{B}_i \neq \emptyset$ and $\mathcal{C}_i$ are pairwise disjoint subsets of the variables, then

$$\mathcal{Q}_k = \cap_{i=1}^{k} \{P \in \mathcal{P} : \mathcal{A}_i \perp\!\!\!\perp \mathcal{B}_i \,|\, \mathcal{C}_i(P)\} \tag{2}$$

is a conditional independence model. In this section we study the properties of $\mathcal{Q}_k$ using the marginal log-linear model framework of BR.

Out of the long tradition of marginal modeling (see, e.g., Prentice & Pyke, 1979, Molenbergs & Verbeke, 2003, Fitzmaurice, Laird, & Ware, 2005 and the references in BR), our approach relies on marginal log-linear models as defined by BR. Marginal models impose restrictions on certain log-linear parameters defined in marginal distributions. To describe precisely which log-linear parameters may be restricted, we need the following notations.

The joint sample space of variables $\mathcal{V}$ is called a contingency table. The joint sample space of a subset of $\mathcal{V}$, say $\mathcal{M}$, is a marginal of the contingency table and the subset itself is also called a marginal. Let $\mathcal{M}_i$, $i = 1, \ldots, m$ be a so-called complete hierarchical order of subsets of $\mathcal{V}$, defined by the property that $\mathcal{M}_i \subseteq \mathcal{M}_j$ implies that $i < j$ and $\mathcal{M}_m = \mathcal{V}$. For every subset $\mathcal{E}$ of $\mathcal{V}$, $\mathcal{M}(\mathcal{E})$ denotes the first marginal in the hierarchical order that contains $\mathcal{E}$. Consider now for all subsets $\mathcal{E}$ the log-linear parameters (see, e.g. Bishop, Fienberg, & Holland, 1975 or Agresti, 2002) within the marginal $\mathcal{M}(\mathcal{E})$. The values of these parameters are associated with different combinations of the indices of the variables in $\mathcal{E}$. Denote certain linearly independent components of this parameter by $\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}$. These are the hierarchical marginal log-linear parameters and saying that a marginal log-linear parameter is zero means that all its components are equal to zero. Hierarchical marginal log-linear models are obtained by assuming that some of

the marginal log-linear parameters are zero. Marginal log-linear parameters and models were systematically studied by BR, see also Bergsma and Rudas (2003) and Bergsma and Rudas (2002b). Note, that in general marginal log-linear models do not have a unique parameterization, since depending on the choice of the marginals, different parameterizations of various models are obtained. Further, a combinatorial property of the marginals, called ordered decomposability, implies variation independence of the parameters. As illustrated by the next example, certain patterns of zeros among the marginal log-linear parameters are equivalent to certain conditional independencies between (groups of) variables:

Example 1 Model $\mathcal{Q}_1$ defined by (1) is a hierarchical marginal log-linear model, which can be seen as follows. Firstly, with $\mathbb{P}(.)$ denoting the power set,

$$\mathbb{D}(\mathcal{A},\mathcal{B},\mathcal{C}) = \mathbb{P}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}) \setminus (\mathbb{P}(\mathcal{A}\cup\mathcal{C}) \cup \mathbb{P}(\mathcal{B}\cup\mathcal{C}))$$

and $\mathcal{M} = \mathcal{A}\cup\mathcal{B}\cup\mathcal{C}$, a well-known fact is that $P \in \mathcal{Q}_1$ if and only if

$$\lambda_{\mathcal{E}}^{\mathcal{M}}(P) = 0 \qquad \forall \mathcal{E} \in \mathbb{D}(\mathcal{A},\mathcal{B},\mathcal{C}),$$

see also Lemma 1, below, and Example 2 for a concrete illustration. Thus, $\mathcal{Q}_1$ is obtained by linear restrictions on the marginal log-linear parameterization generated by the hierarchical and complete list of marginals $(\mathcal{M},\mathcal{V})$, and hence is a hierarchical marginal log-linear model. □

An important aspect of a model is smoothness. A model is said to be smooth if it admits a smooth parameterization. A function of the probability distributions in the model is called a parameter and it is a parameterization if it is invertible. The inverse is called the probability function. A parameterization is smooth, if the probability function is twice continuously differentiable with a full rank Jacobian and the model, viewed as a set of probability distributions, is a the image of an open set in a Euclidean space. See BR for details. BR proved that, for a complete hierarchical order $\mathcal{M}_1,\ldots,\mathcal{M}_m$,

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} : \ \mathcal{E} \subseteq \mathcal{V}\}$$

is a smooth parameterization of the joint distribution of $\mathcal{V}$. It follows that hierarchical marginal log-linear models are smooth.

When a model is defined by certain restrictions on the probability distributions, such a specification is called minimal if no restrictions can be removed without changing the model. This was first studied in the context

of certain marginal log-linear models by Lang and Agresti (1994), see also Bergsma et al. (2009).

The next theorem applies the above framework to the conditional independence model $\mathcal{Q}_k$. Let

$$\mathbb{D}_i = \mathbb{D}(\mathcal{A}_i, \mathcal{B}_i, \mathcal{C}_i), \qquad i = 1, \ldots, k.$$

Theorem 1 *Suppose there exists a sequence $\mathcal{M}_1, \ldots, \mathcal{M}_m$ of subsets of $\mathcal{V}$ in complete hierarchical order that satisfies*

$$\mathcal{C}_i \subseteq \mathcal{M}(\mathcal{E}) \subseteq \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i \qquad \forall i \leq k, \mathcal{E} \in \mathbb{D}_i\,. \tag{3}$$

Then the following statements hold true:

S1: A distribution Q is in $\mathcal{Q}_k$ if and only if

$$\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}(Q) = 0 \qquad \forall \mathcal{E} \in \cup_{i=1}^{k} \mathbb{D}_i. \tag{4}$$

S2: $\mathcal{Q}_k$ is a hierarchical marginal log-linear model and is hence smooth.

S3: $\mathcal{Q}_k$ is parameterized by

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} \,:\, \mathcal{E} \notin \cup_{i=1}^{k} \mathbb{D}_i\} \tag{5}$$

and this is a smooth parameterization.

S4: The specification of $\mathcal{Q}_k$ given in (4) *is minimal.*

The proof of the theorem is postponed until the end of this section. Note that Theorem 1 implies that the number of degrees of freedom associated with $\mathcal{Q}_k$ is

$$\sum_{\mathcal{E} \in \cup_{i=1}^{k} \mathbb{D}_i} \prod_{V \in \mathcal{E}} (C_V - 1)\ ,$$

where C_V is the number of categories of variable V.

Example 2 below illustrates the theorem and Example 3 shows a limitation of it.

Example 2 For the models $A \perp\!\!\!\perp B \,|\, C$ and $A \perp\!\!\!\perp C \,|\, D$, $\mathbb{D}_1 = \{AB, ABC\}$ and $\mathbb{D}_2 = \{AC, ACD\}$. The sequence of marginals $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3) = (ABC, ACD, ABCD)$ is such that (3) is satisfied and the model is hierarchical marginal log-linear.

For the intersection of $A \perp\!\!\!\perp B|C$ and $AD \perp\!\!\!\perp C|B$, $\mathbb{D}_1 = \{AB, ABC\}$ and $\mathbb{D}_2 = \{AC, CD, ACD, ABC, BCD, ABCD\}$. Then, with $(\mathcal{M}_1, \mathcal{M}_2) = (ABC, ABCD)$, condition (3) is satisfied, so the model is hierarchical marginal log-linear and smooth.

A minimal specification for this model is

$$\lambda_{AB}^{ABC} = 0 \qquad \lambda_{AC}^{ABC} = 0 \qquad \lambda_{ABC}^{ABC} = 0 \tag{6}$$

$$\lambda_{CD}^{ABCD} = 0 \qquad \lambda_{ACD}^{ABCD} = 0 \qquad \lambda_{BCD}^{ABCD} = 0 \qquad \lambda_{ABCD}^{ABCD} = 0 \tag{7}$$

By calculating the number of linearly independent restrictions for each parameter, the number of degrees of freedom of the model may be determined.

It might be thought that the additional restriction

$$\lambda_{AC}^{ABCD} = 0 \qquad \lambda_{ABC}^{ABCD} = 0 \tag{8}$$

is needed to specify the model, since (7) and (8) together are equivalent to $AD \perp\!\!\!\perp C|B$ (see Lemma 1 and Example 1). However, it follows from Theorem 1 that (8) is implied by (6) and (7). Thus, application of Theorem 1 is necessary to achieve minimal specification of the model. □

Example 3 The model defined as the intersection of $A \perp\!\!\!\perp B|D$, $A \perp\!\!\!\perp C|B$, $A \perp\!\!\!\perp D|C$ is not identified as a smooth model by Theorem 1, although from the inspection of the Jacobian we suspect that it is, in fact, smooth. □

The constraints of a nonsmooth model have a Jacobian with nonconstant rank. A relevant result was obtained by BR who showed that for $\mathcal{M} \neq \mathcal{N}$, the derivatives

$$\frac{d\lambda_{\mathcal{E}}^{\mathcal{M}}(P)}{dP} \text{ and } \frac{d\lambda_{\mathcal{E}}^{\mathcal{N}}(P)}{dP}$$

are equal when evaluated at the uniform distribution, but unequal at some other distributions. Thus, the model defined by setting $\lambda_{\mathcal{E}}^{\mathcal{M}} = \lambda_{\mathcal{E}}^{\mathcal{N}} = 0$ is nonsmooth. However, if there are additional restrictions in the model specification, the model may be smooth, as was seen in the second example in Example 2, where the $\mathbb{D}_i$ are not disjoint but the model is still smooth.

If some of the $\mathbb{D}_i$'s are not disjoint (the same effect is restricted in at least two different marginal tables), we have so far always found that the specification is not minimal and/or the model is nonsmooth. This is illustrated in Example 4.

Example 4 The intersection of $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \,|\, C$, discussed earlier, is equivalent to the union of $A \perp\!\!\!\perp BC$ and $B \perp\!\!\!\perp AC$, which is not smooth at the points where the three variables are mutually independent (Dawid, 1980, Example 7 in BR). It can be verified that the condition of the theorem is not fullfilled. Here, $\mathbb{D}(\mathcal{A}_1, \mathcal{B}_1, \mathcal{C}_1) = \{AB\}$ and $\mathbb{D}(\mathcal{A}_2, \mathcal{B}_2, \mathcal{C}_2) = \{AB, ABC\}$, which have the nonempty intersection $\{AB\}$.

Drton (2009) gives two examples of nonsmooth models: (i) $B \perp\!\!\!\perp D|AC$ and $A \perp\!\!\!\perp BD$, (ii) $B \perp\!\!\!\perp D|A$ and $A \perp\!\!\!\perp BD|C$. For both the condition of the theorem is not satisfied, and the $\mathbb{D}_i$s are not disjoint. □

We conclude this section with the proof of Theorem 1. In the proof we use the next lemma which summarizes well-known properties of conditional independence models:

Lemma 1 *Let $P \in \mathcal{P}$ and let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be pairwise disjoint subsets of $\mathcal{V}$. Then, the following four properties are equivalent*

1. $\mathcal{A} \perp\!\!\!\perp \mathcal{B}|\mathcal{C}(P)$
2. $P(\mathcal{ABC}) = \frac{P(\mathcal{AC})P(\mathcal{BC})}{P(\mathcal{C})}$
3. $\lambda_{\mathcal{D}}^{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(P) = 0 \quad \forall \mathcal{D} \in \mathbb{D}(\mathcal{A}, \mathcal{B}, \mathcal{C})$
4. $P(\mathcal{ABC}) = t(\mathcal{AC})u(\mathcal{BC})$ *for some functions t and u.*

Proof of Theorem 1: First we prove is that S1 is true. We start by showing that $Q \in \mathcal{Q}_k$ implies (4).

Let $i \in \{1, \ldots, k\}$ and $\mathcal{E} \in \mathbb{D}_i$ be arbitrary. If $Q \in \mathcal{Q}_k$, then

$$\mathcal{A}_i \cap \mathcal{M}(\mathcal{E}) \perp\!\!\!\perp \mathcal{B}_i \cap \mathcal{M}(\mathcal{E}) \,|\, \mathcal{C}_i(Q).$$

Now by the first and second inclusion of (3), respectively,

$$\begin{aligned}
[\mathcal{A}_i \cap \mathcal{M}(\mathcal{E})] \cup [\mathcal{B}_i \cap \mathcal{M}(\mathcal{E})] \cup \mathcal{C}_i &= [\mathcal{A}_i \cap \mathcal{M}(\mathcal{E})] \cup [\mathcal{B}_i \cap \mathcal{M}(\mathcal{E})] \cup [\mathcal{C}_i \cap \mathcal{M}(\mathcal{E})] \\
&= [\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i] \cap \mathcal{M}(\mathcal{E}) \\
&= \mathcal{M}(\mathcal{E})
\end{aligned}$$

In addition, it is easily verified that $\mathcal{E} \in \mathbb{D}(\mathcal{A}_i \cap \mathcal{M}(\mathcal{E}), \mathcal{B}_i \cap \mathcal{M}(\mathcal{E}), \mathcal{C}_i)$, so Lemma 1 implies

$$\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}(Q) = 0$$

Since $i \in \{1, \ldots, k\}$ and $\mathcal{E} \in \mathbb{D}_i$ were chosen arbitrarily, (4) follows.

It remains to be proven that if (3) and (4) hold, then $Q \in \mathcal{Q}_k$. For $j \leq m$, the index set

$$\mathcal{I}_j = \{i \leq k | \mathcal{C}_i \subseteq \mathcal{M}_j \subseteq \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i\}$$

refers to those conditional independencies in the specification of $\mathcal{Q}_k$ which imply a conditional independence restriction for marginal $\mathcal{M}_j$: with $\mathcal{A}_{ij} = \mathcal{A}_i \cap \mathcal{M}_j$ and $\mathcal{B}_{ij} = \mathcal{B}_i \cap \mathcal{M}_j$, it easily follows that

$$\mathcal{M}_j = \mathcal{A}_{ij} \cup \mathcal{B}_{ij} \cup \mathcal{C}_i \qquad \forall i \in \mathcal{I}_j$$

and the conditional independencies for $\mathcal{M}_j$ are

$$\mathcal{A}_{ij} \perp\!\!\!\perp \mathcal{B}_{ij} | \mathcal{C}_i(Q) \qquad \forall i \in \mathcal{I}_j. \tag{9}$$

Since $\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i \in \mathbb{D}_i$, by (3) we must have that for all $i \leq k$ there is a $j_i \leq m$ such that $\mathcal{M}_{j_i} = \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i$. But then $\mathcal{A}_{ij_i} = \mathcal{A}_i$ and $\mathcal{B}_{ij_i} = \mathcal{B}_i$, so choosing $j = j_i$ in (9) shows that if (9) holds for all $j \leq m$, then $Q \in \mathcal{Q}_k$. To complete the proof of *2.*, it is therefore sufficient to show that (4) implies that (9) holds for all $j \leq m$.

Suppose a distribution Q on the contingency table satisfies (4). Let $\mathcal{M}_0 = \emptyset$. We now show Q also satisfies (9) for all $j = 0, 1, \ldots, m$ by induction on j. For $j = 0$, (9) is trivially satisfied. For the induction step, assume that (9) is true for $j = 0, \ldots, l-1$ for some $l \leq m$. We show that (9) holds for $j = l$. For this we construct a sequence of distributions $Q_0, Q_1, \ldots$ by means of the iterative proportional fitting procedure (IPFP, see Bishop et al., 1975), converging to a distribution Q^*, such that each distribution in the sequence satisfies (9) for $j = l$, and so does Q^* and show that $Q^* = Q$, thus completing the proof of *2.*

With

$$\mathbb{E}_l = \mathbb{P}(\mathcal{M}_l) \cap (\cup_{i \in \mathcal{I}_l} \mathbb{D}_i)$$

we define Q_0 on $\mathcal{M}_l$ by its log-linear parameterization,

$$\lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q_0) = 0 \qquad \forall \mathcal{E} \in \mathbb{E}_l \tag{10}$$

and

$$\lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q_0) = \lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q) \qquad \forall \mathcal{E} \in \mathbb{P}(\mathcal{M}_l) \setminus \mathbb{E}_l \tag{11}$$

Then define Q_n $(n = 1, 2, \ldots)$ on $\mathcal{M}_l$ by an IPFP step as

$$Q_n(\mathcal{M}_l) = Q_{n-1}(\mathcal{M}_l) \frac{Q(\mathcal{M}_j \cap \mathcal{M}_l)}{Q_{n-1}(\mathcal{M}_j \cap \mathcal{M}_l)}, \tag{12}$$

where $j = n \bmod m$ if n is not a multiple of m, and $j = m$ otherwise.

We now show by (an inner) induction on n that (9) with $j = l$ holds for Q_n. For Q_0 this is immediate by (10) and Lemma 1. The inner induction assumption is that (9) holds, instead of Q, for Q_{n-1} ($n \geq 1$), for $j = l$. By Lemma 1, (9) for a given $j \leq m$ is equivalent to the existence of a factorization

$$Q(\mathcal{A}_{ij} \cup \mathcal{B}_{ij} \cup \mathcal{C}_i) = t_{ij}(\mathcal{A}_{ij} \cup \mathcal{C}_i)u_{ij}(\mathcal{B}_{ij} \cup \mathcal{C}_i) \qquad \forall i \in \mathcal{I}_j \tag{13}$$

for certain functions t and u. Clearly, with $\mathcal{A}_{ijl} = \mathcal{A}_{ij} \cap \mathcal{M}_l$ and $\mathcal{B}_{ijl} = \mathcal{B}_{ij} \cap \mathcal{M}_l$, it follows that also

$$Q(\mathcal{A}_{ijl} \cup \mathcal{B}_{ijl} \cup \mathcal{C}_i) = t_{ijl}(\mathcal{A}_{ijl} \cup \mathcal{C}_i)u_{ijl}(\mathcal{B}_{ijl} \cup \mathcal{C}_i) \qquad \forall i \in \mathcal{I}_j \tag{14}$$

for certain functions t and u. By the inner induction assumption, factorization (13) with $j = l$ holds for $Q_{n-1}(\mathcal{M}_l)$ and factorization (14) holds for $Q_{n-1}(\mathcal{M}_j \cap \mathcal{M}_l)$. Since $Q(\mathcal{M}_j \cap \mathcal{M}_l)$ also factorizes as in (14), by (12) Q_n also factorizes as required, completing the inner induction step.

Since the marginal distributions $Q(\mathcal{M}_j \cap \mathcal{M}_l)$ are consistent, because they are all marginal distributions of Q, and since Q is assumed strictly positive, Q_n converges uniformly (Csiszár, 1975) to a distribution on $\mathcal{M}_l$, say Q^*. Because of uniform convergence, Q^* also satisfies (9) for $j = l$.

We now complete the induction step by showing that $Q^*(\mathcal{M}_l) = Q(\mathcal{M}_l)$. The distribution Q restricted to $\mathcal{M}_l$ is characterized by a mixed parameterization of it in the exponential family sense (Barndorff-Nielsen (1978), see also Rudas (1998)), using the marginal distributions

$$\{Q(\mathcal{M}_j \cap \mathcal{M}_l) | j \leq m\} \tag{15}$$

and the log-linear parameters

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q) | \mathcal{E} \in \mathbb{P}(\mathcal{M}_l) \setminus \cup_{j=1}^{l-1} \mathbb{P}(\mathcal{M}_j \cap \mathcal{M}_l)\} \tag{16}$$

Since Q^* is the fixed point solution of the IPFP, Q^* has the marginals given in (15). Furthermore, the parameters

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q^*) | \mathcal{E} \in \mathbb{P}(\mathcal{M}_l) \setminus \cup_{j=1}^{l-1} \mathbb{P}(\mathcal{M}_j \cap \mathcal{M}_l)\} \tag{17}$$

are equal to the corresponding parameters for Q_0, because these are left unchanged by the IPFP, and by (10) and (11) these are the same as (16). Hence, $Q^*(\mathcal{M}_l) = Q(\mathcal{M}_l)$ and since $Q^*(\mathcal{M}_l)$ satisfies (9) for $j = l$, so does $Q(\mathcal{M}_l)$, completing the induction step, and thereby the proof of S1.

To see the rest of the Theorem, note that, since $\mathcal{M}_1, \ldots, \mathcal{M}_m$ is hierarchical and complete, Theorem 2 of BR can be applied with, using the notation of that paper, $\mathcal{P}$ and $\tilde{\lambda}_{\mathcal{P}}$ defined as $\mathcal{P} = \{(\mathcal{E}, \mathcal{M}(\mathcal{E})) \,|\, \mathcal{E} \subseteq \mathcal{V}\}$ and $\tilde{\lambda}_{\mathcal{P}} = \{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} \,|\, \mathcal{E} \subseteq \mathcal{V}\}$, implying that the latter is a smooth parameterization of the distributions on the contingency table. This and S1 then immediately imply S3.

Theorem 5 in BR can now be applied as well and, by S1, $\mathcal{Q}_k$ is a hierarchical marginal loglinear model and curved exponential, hence it is smooth, i.e., S2 holds.

Finally, since $\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} \,|\, \mathcal{E} \subseteq \mathcal{V}\}$ is a smooth parameterization, it is one-to-one and S4 follows. □

3 Block recursive Markov models

This section investigates certain chain graph models from the perspective of marginal models. Since the seminal paper of Darroch, Lauritzen, and Speed (1980) was published, several articles proposed Markovian statistical models associated with various graphs, including chain graphs containing various combinations of (undirected) lines and (directed) arrows of solid or dashed types (Lauritzen & Wermuth, 1989, Frydenberg, 1990, Cox & Wermuth, 1996, Andersson et al., 2001, Richardson, 2003, Wermuth & Cox, 2004). These statistical models assume certain conditional or marginal independence properties of the joint distribution of the variables that are identified with the nodes of the graph.

A chain graph $\mathcal{G}$ is a graph with lines and arrows that has no loops and semi-directed cycles. An undirected path between two nodes is a path consisting of lines only. The equivalence classes of nodes that are connected by an undirected path are called the components of $\mathcal{G}$. Note that this is the definition of chain graphs used in, e.g., Andersson et al. (2001). There is a related but different definition used by, e.g., Cox and Wermuth (1996) in which the components are completely ordered but not necessarily connected. In the definition used here, the components are partially ordered by the direction of the arrows because there are no semi-directed cycles. Chain graphs are also called joint response graphs (Cox & Wermuth, 1996) and may be used to model joint distributions where some of the variables (those within a component) are jointly responses to explanatory variables (in earlier components). A line within a component represents some kind of association between variables and an arrow represents some kind of direct effect from

an earlier variable to a later variable. As Andersson et al. (2001) say it, a chain graph represents structural and associative dependencies.

The following notation is used. For a component $\mathcal{K}$, $\mathrm{ND}(\mathcal{K})$ is the set of nondescendants of $\mathcal{K}$, i.e., the union of those components, except $\mathcal{K}$, for which no semi-directed path leads from any node in $\mathcal{K}$ to any node in these components. $\mathrm{PA}(\mathcal{K})$ is the set of parents of $\mathcal{K}$, i.e., the union of those components from which an arrow points to a node in $\mathcal{K}$. The set of neighbors of $\mathcal{X} \subseteq \mathcal{K}$, $\mathrm{nb}(\mathcal{X})$, is the set of nodes in $\mathcal{K}$ that are connected (by a line) to any node in $\mathcal{X}$ and $\mathrm{pa}(\mathcal{X})$ is the set of nodes from which an arrow points to any node in $\mathcal{X}$.

In defining the block-recursive versions of Markov chain graph models we follow Drton (2009) in considering four variants of such models using the definitions given in Andersson et al. (2001). These models are defined by three of the following properties given below: Property 1 and either Property 2a or Property 2b, and either Property 3a or Property 3b.

Property 1: For any component $\mathcal{K}$,

$$\mathcal{K} \perp\!\!\!\perp \{\mathrm{ND}(\mathcal{K}) \setminus \mathrm{PA}(\mathcal{K})\} \,|\mathrm{PA}(\mathcal{K}),$$

Property 2a: For all subsets $\mathcal{X} \subseteq \mathcal{K}$,

$$\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X})\} \,|\, \{\mathrm{PA}(\mathcal{K}) \cup \mathrm{nb}(\mathcal{X})\},$$

Property 2b: For all subsets $\mathcal{X} \subseteq \mathcal{K}$,

$$\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X})\} \,|\mathrm{PA}(\mathcal{K}).$$

Property 3a: For all components $\mathcal{K}$ and subsets $\mathcal{X} \subseteq \mathcal{K}$,

$$\mathcal{X} \perp\!\!\!\perp \{\mathrm{PA}(\mathcal{K}) \setminus \mathrm{pa}(\mathcal{X})\} \,|\, \{\mathrm{pa}(\mathcal{X}) \cup \mathrm{nb}(\mathcal{X})\},$$

Property 3b: For all components $\mathcal{K}$ and subsets $\mathcal{X} \subseteq \mathcal{K}$,

$$\mathcal{X} \perp\!\!\!\perp \{\mathrm{PA}(\mathcal{K}) \setminus \mathrm{pa}(\mathcal{X})\} \,|\mathrm{pa}(\mathcal{X}).$$

The following four Markov properties of chain graphs were considered by Drton (2009):

Type I: Properties 1, 2a and 3a

Type II: Properties 1, 2a and 3b

Type III: Properties 1, 2b and 3a

Type IV: Properties 1, 2b and 3b

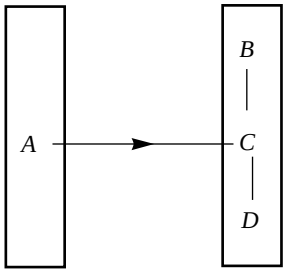

Figure 1: Chain graph leading to a nonsmooth model under the Type II and Type III Markov properties

The Type I Markov property is also called the LWF block-recursive Markov property (see Lauritzen & Wermuth, 1989 and Frydenberg, 1990) and the Type II Markov property is also called the AMP block-recursive Markov property (see Andersson et al., 2001).

In the next example a particular chain graph is considered, which under Types II and III Markov properties gives a nonsmooth model, and under Types I and IV gives a smooth model.

Example 5 Consider a chain graph with two components, the first one containing variable A and the second one containing variables B, C, D. An arrow points from A to C and lines connect B with C and C with D, as shown in Figure 1. Property 1 has no conditional independence implications, Property 2a implies

$$B \perp\!\!\!\perp D \,|\, AC \ ,$$

Property 2b implies

$$B \perp\!\!\!\perp D \,|\, A \ ,$$

Property 3a implies

$$A \perp\!\!\!\perp BD \,|\, C \ ,$$

and Property 3b implies

$$A \perp\!\!\!\perp BD \ .$$

Drton (2009) used algebraic methods to show that both the Type II and Type III models are nonsmooth. This was pointed out to us by two anonymous reviewers. In both cases, it can be verified that the conditions of Theorem 1 are violated.

The Type I and Type IV models are standard log-linear models for this graph and hence smooth. Note that in these cases the conditions of Theorem 1 are satisfied. □

To see that Type I models are smooth, consider the components of the chain graph in a well-numbering (see Lauritzen, Dawid, Larsen, & Leimer, 1990) $\mathcal{K}_t$, $t = 1, \ldots, u$. In a well-numbering, parents of a node always precede that node. Note that the proof Lauritzen et al. (1990) for the existence of a well-numbering applies to the nodes of a directed acyclic graph (DAG) but the components of a chain graph may be identified with the nodes of a DAG.

Frydenberg (1990) proved that a positive distribution belongs to the Type I model, if and only if (i) the distribution factorizes into conditional distributions $P(\mathcal{K}|\mathrm{pa}(\mathcal{K}))$ of components given the parents of their nodes, and (ii) the undirected pairwise Markov property holds for the joint distributions of $\mathcal{K}\cup\mathrm{pa}(\mathcal{K})$ with respect to the moral graph of $\mathcal{K}\cup\mathrm{pa}(\mathcal{K})$, $(\mathcal{K}\cup\mathrm{pa}(\mathcal{K}))_m$, that is defined by replacing arrows with lines in the subgraph $\mathcal{K} \cup \mathrm{pa}(\mathcal{K})_c$, where $\mathrm{pa}(\mathcal{K})_c$ is a complete graph on the nodes in $\mathrm{pa}(\mathcal{K})$. Condition (i) is equivalent, for all t, to

$$\mathcal{K}_t \perp\!\!\!\perp \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{t-1} \setminus \mathrm{pa}(\mathcal{K}_t) | \mathrm{pa}(\mathcal{K}_t),$$

and these conditional independencies may be parameterized in marginals of the form

$$\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t,$$

where the effects that have to have zero log-linear parameters are those in $\mathbb{D}(\mathcal{K}_t, \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{t-1} \setminus \mathrm{pa}(\mathcal{K}_t), \mathrm{pa}(\mathcal{K}_t))$. Condition (ii) may be parameterized, for all t, in marginals of the form

$$\mathcal{K}_t \cup \mathrm{pa}(\mathcal{K}_t)$$

and the effects that have zero log-linear parameters are those containing two variables from $(\mathcal{K}_t \cup \mathrm{pa}(\mathcal{K}_t))_m$ that are not connected by a line. Further, for any $\mathcal{E} \in \mathbb{D}(\mathcal{K}_t, \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{t-1} \setminus \mathrm{pa}(\mathcal{K}_t), \mathrm{pa}(\mathcal{K}_t))$, $\mathcal{E}$ is neither a subset of $\mathcal{K}_v \cup \mathrm{pa}(\mathcal{K}_v)$ for any $v \leq t$ nor of $\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_v$, for any $v < t$. Also, if $\mathcal{E} \in \mathcal{K}_t \cup \mathrm{pa}(\mathcal{K}_t)$ and $\mathcal{E}$ contains two variables that are not connected in

$(\mathcal{K}_t \cup \mathrm{pa}(\mathcal{K}_t))_m$, then $\mathcal{E}$ is neither a subset of $\mathcal{K}_v \cup \mathrm{pa}(\mathcal{K}_v)$ nor of $\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_v$, for any $v < t$, because at least one of these two variables is in $\mathcal{K}_t$.

Thus a marginal log-linear parameterization of distributions in a Type I model may be obtained using pairs of subsets of the form $\mathcal{K}_t \cup \mathrm{pa}(\mathcal{K}_t), \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t$, for all components t. The list of marginals

$$\mathcal{K}_1 \cup \mathrm{pa}(\mathcal{K}_1), \mathcal{K}_1, \mathcal{K}_2 \cup \mathrm{pa}(\mathcal{K}_2), \mathcal{K}_1 \cup \mathcal{K}_2, \ldots, \mathcal{K}_u \cup \mathrm{pa}(\mathcal{K}_u), \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_u$$

is in a hierarchical order, implying that distributions in Type I models are smooth.

Turning to Type IV models, Drton (2009) showed that for any chain graph, these models are smooth. We give an alternative proof of smoothness and also provide a parameterization.

Theorem 2 *Assuming strictly positive discrete distributions, a Type IV model for a chain graph is a hierarchical marginal log-linear model, and is, therefore, smooth. The parameterization is based on the marginals*

$$\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t$$

and the parameters set to zero are those associated with effects in

$$\{\mathbb{ID}(\mathcal{X}, \mathcal{K}_t \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X}), \mathrm{PA}(\mathcal{K}_t)) : \mathcal{X} \subseteq \mathcal{K}\} \cup \tag{18}$$

$$\{\mathbb{ID}(\mathcal{X}, \mathrm{PA}(\mathcal{K}) \setminus \mathrm{pa}(\mathcal{X}), \mathrm{pa}(\mathcal{X})) : \mathcal{X} \subseteq \mathcal{K}\} \cup$$

$$\mathbb{ID}(\mathcal{K}_t, \mathrm{ND}_p(\mathcal{K}_t) \setminus \mathrm{PA}(\mathcal{K}_t), \mathrm{PA}(\mathcal{K}_t)).$$

Proof: For each component $\mathcal{K}_t$, the conditioning set in 2b is $\mathrm{PA}(\mathcal{K}_t)$ and in 3b it is $\mathrm{pa}(\mathcal{X}) \subseteq \mathrm{PA}(\mathcal{K}_t)$, thus for all conditional independencies implied by 2b or 3b, if written in the form of $\mathcal{A}_i \perp\!\!\!\perp \mathcal{B}_i | \mathcal{C}_i$, $\mathcal{C}_i \subseteq \mathrm{PA}(\mathcal{K}_t)$. Further, for these conditional independencies, $\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i \subseteq \mathcal{K}_t \cup \mathrm{PA}(\mathcal{K}_t)$. Thus, for a hierarchical order of the marginals

$$\{PA(\mathcal{K}_t) \cup \mathcal{X} : \mathcal{X} \subseteq \mathcal{K}_t\} \tag{19}$$

condition (3) holds and Theorem 1 applies. Therefore, a hierarchical marginal log-linear parameterization of the distribution with Properties 2b and 3b for any $\mathcal{K}_t$ is obtained.

In addition to the conditional independencies in 2b and 3b, the independency in Property 1 has to be imposed.

Lauritzen et al. (1990) defined a Markov property (MP) called local well-numering MP for DAGs that is based on a well-numbering on the nodes of

the DAG. Their Theorem 5 is that for a distribution on a DAG, the local well-numbering MP holds if and only if the local directed MP does. Their proof applies to components of chain graphs, so it is also true that for a distribution on the chain graph, Property 1 holds if and only if the following Property 4 does.

Property 4: For any component $\mathcal{K}_t$,

$$\mathcal{K}_t \perp\!\!\!\perp \{\mathrm{ND}_p(\mathcal{K}_t) \setminus \mathrm{PA}(\mathcal{K}_t)\} \,|\mathrm{PA}(\mathcal{K}_t),$$

where $\mathrm{ND}_p(\mathcal{K}_t)$ is the set of those nondescendants of $\mathcal{K}_t$ that precede it in the ordering. It is easy to see that $\mathrm{ND}_p(\mathcal{K}_t) = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{t-1}$. Property 4 may be parameterized using the marginal

$$\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t. \tag{20}$$

Including the marginal in (20) after the marginals in (19) is hierarchical and the conditional independency in Property 4 is parameterized by setting to zero marginal log-linear parameters (see Lemma 1) that are associated with effects appearing for the first time in (20).

Hence a hierarchical marginal log-linear parameterization of the Type IV block recursive model may be obtained by using the marginals in (19) and (20):

$$\{PA(\mathcal{K}_t) \cup \mathcal{X} : \mathcal{X} \subseteq \mathcal{K}_t\}^*, \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t, \tag{21}$$

where $\{\ \}^*$ denotes a hierarchical ordering of the elements of the set.

Theorem 1 implies that in the parameterization based on the marginals in (21) the effects associated with (18) are zero and the remaining parameters parameterize the distributions in the model. □

As a closing remark, we note that using the results of this section, path models based on block recursive chain graph models could be defined and studied, just as it was done with path models based on directed acyclic graphs in Rudas, Bergsma, and Németh (2006), but this is not going to be pursued here.

4 Acknowledgements

Part of the research of the first author was done while he was a visitor in the Center for Statistics & the Social Sciences and the Department of Statistics, University of Washington, Seattle, where he is now an Affiliate Professor. The first author is also a Recurrent Visiting Professor with the Central

European University, Budapest, and the moral support received is acknowledged. Part of the second author's research was done while he was working at EURANDOM in Eindhoven, the Netherlands, and was additionally supported by PASCAL travel grants. The authors thank Thomas Richardson and Michael Perlman for several discussions and two anonymous reviewers for helpful comments that greatly influenced the contents of the paper and, in particular, for drawing our attention to the related results in algebraic statistics.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd edition.* New York: Wiley.

Andersson, S. A., Madigan, D., & Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, *28*, 33-85.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families.* New York: Wiley.

Bergsma, W. P. (1997). *Marginal models for categorical data.* Tilburg: Tilburg University Press.

Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal Models for Dependent, Clustered and Longitudinal Categorical Data.* New York: Springer.

Bergsma, W. P., & Rapcsak, T. (2006). An exact penalty method for smooth equality constrained optimization with application to maximum likelihood estimation. *EURANDOM-report 2006-001*.

Bergsma, W. P., & Rudas, T. (2002a). Marginal models for categorical data. *Annals of Statistics*, *30*, 140-159.

Bergsma, W. P., & Rudas, T. (2002b). Variation independent parameterizations of categorical distributions. *In: Distributions with Given Marginals and Statistical Modelling, eds. C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena. Kluwer.*, 21–27.

Bergsma, W. P., & Rudas, T. (2003). On conditional and marginal association. *Annales de la Faculte des Sciences de Toulouse*, *11*, 455-468.

Bishop, Y. V. V., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

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

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, *3*, 146-158.

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

Dawid, A. P. (1980). Conditional independence for statistical operations. *Ann. Stat.*, *8*, 598-617.

Drton, M. (2009). Discrete chain graph models. *Bernoulli*, *Forthcoming*.

Drton, M., Sturmfels, B., & Sullivant, S. (2009). *Lectures on algebraic statistics*. Basel: Birkhauser.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2005). *Applied Longitudinal Analysis*. Wiley.

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, *17*, 333-353.

Glonek, G. J. N., & McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B*, *57*, 533-546.

Lang, J. B., & Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *J. Am. Stat. Ass.*, *89*, 625-632.

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

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., & Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, *20*, 491-505.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, *17*, 31-57.

Molenbergs, G., & Verbeke, G. (2003). *Models for Discrete Longitudinal Data*. Springer.

Prentice, R. L., & Pyke, R. (1979). Logistic incidence models and case-control studies. *Biometrika*, *60*, 403-411.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, *30*, 145-157.

Rudas, T. (1998). *Odds Ratios in the Analysis of Contingency Tables*. Newbury Park: Sage.

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

Studeny, M. (2004). *Probabilistic conditional independence structures*. New York: Springer.

Wermuth, N., & Cox, D. R. (2004). Joint response graphs and separation

induced by triangular systems. *Journal of the Royal Statistical Society B*, *66*, 687-717.