# Marginal log-linear parameterization of conditional independence models

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

March 21, 2010

**Abstract**

Models defined by a set of conditional independence restrictions play an important role in statistical theory and applications, especially, but not only, in graphical modeling. In this paper we identify a subclass of these consisting of hierarchical marginal log-linear models, 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 of chi-squared statistics to test goodness-of-fit. The utility of the results is illustrated by applying them to certain block-recursive Markov models associated with chain graphs.

Key words: Chain graph; Conditional independence; Graphical model; Marginal model; Smoothness.

# 1 Introduction

Conditional independence models have received considerable attention recently, see, e.g., Studeny (2005). Such models are defined by one or more conditional independence restrictions on a set of random variables. Graphical models are perhaps the most important examples, see Cox and Wermuth (1996), Lauritzen (1996) and the references in Section 3.

Conditional independence models may show unexpected behaviour. For example, for random variables $A$, $B$, and $C$, the intersection of $A \perp\!\!\!\perp C$ and $A \perp\!\!\!\perp B \mid C$ can be verified to be equivalent to $A \perp\!\!\!\perp BC$, where $BC$ means the joint distribution of $B$ and $C$. But if $C$ is dichotomous, the intersection of $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \mid C$ is equivalent to the union of $A \perp\!\!\!\perp BC$ and $B \perp\!\!\!\perp AC$, and has singularities, see Bergsma and Rudas (2002a).

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). Such models are smooth, a characteristic that aids their interpretation, and guarantees the applicability of standard asymptotic theory. The intersection of $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B \mid C$ is singular 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), Bergsma, Croon, and Hagenaars (2009), and allows for easy computation of the degrees of freedom for the model.

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, and Perlman (2001), are smooth. This result is not new but our approach gives an interpretable parameterization and implies the number of degrees of freedom.

# 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 $\mathcal{V}$. Further, for $i = 1, \ldots, k$, let

$\mathcal{A}_i \neq \emptyset$, $\mathcal{B}_i \neq \emptyset$ and $\mathcal{C}_i$ be pairwise disjoint subsets of the variables. Then

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

is a conditional independence model which consists of the probability distributions $P$ for which the required conditional independencies $\mathcal{A}_i \perp\!\!\!\perp \mathcal{B}_i \mid \mathcal{C}_i$, $i = 1, \ldots, k$, hold. In this section properties of $\mathcal{Q}_k$ are studied using the marginal log-linear model framework of Bergsma and Rudas (2002a). Marginal log-linear models impose restrictions on log-linear parameters defined in marginal distributions.

The joint sample space of variables $\mathcal{V}$ is called a contingency table and that of a subset of $\mathcal{V}$, say $\mathcal{M}$, is a marginal of the contingency table. Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ 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 $i < j$ 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 corresponding log-linear parameter (Bishop, Fienberg, and Holland, 1975 or Agresti, 2002) within the marginal $\mathcal{M}(\mathcal{E})$. The values of the components of this parameter are associated with different combinations of the indices of the variables in $\mathcal{E}$. Denote a choice of maximal linearly independent components by $\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}$. These are the hierarchical marginal log-linear parameters. The assumption that $\Lambda = \{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} : \mathcal{E} \subseteq \mathcal{V}\}$ is in a linear subspace of a $|\Lambda|$-dimensional Euclidean space is a hierarchical marginal log-linear model. Marginal log-linear parameters and models were systematically studied by Bergsma and Rudas (2002a), see also Bergsma and Rudas (2002b). In general, marginal log-linear models do not have a unique parameterization, since depending on the choice of the marginals, different parameterizations of the same model are obtained.

A model is a set of probability distributions and an important property of a model is smoothness. A model is 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. A parameterization is smooth, if it is a twice continuously differentiable homeomorphism onto an open set in Euclidean space. Bergsma and Rudas (2002a) 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 $\mathcal{V}$.

For models defined by restrictions on a parameterization, the specification is called minimal if no restriction can be removed without changing the

model. This was first studied in the context of marginal log-linear models by Lang and Agresti (1994).

To apply this framework to $\mathcal{Q}_k$, define, with $\mathbb{P}(.)$ denoting the power set,

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

**Theorem 1** *For the model defined by (1), suppose there exists a sequence $\mathcal{M}_1, \ldots, \mathcal{M}_m$ of subsets of $\mathcal{V}$ in complete hierarchical order that satisfies*

$$\mathcal{C}_i \subseteq \mathcal{M}(\mathcal{E}) \subseteq \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i, \qquad i = 1, \ldots, k,\ \mathcal{E} \in \mathbb{D}_i\,. \tag{2}$$

*Then the following statements hold true:*

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

$$\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}(Q) = 0, \qquad \mathcal{E} \in \cup_{i=1}^{k} \mathbb{D}_i. \tag{3}$$

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

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

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} \,:\, \mathcal{E} \in \mathbb{P}(\mathcal{V}) \setminus \cup_{i=1}^{k} \mathbb{D}_i\} \tag{4}$$

*and this is a smooth parameterization.*

*S4: The specification of $\mathcal{Q}_k$ given in (3) is minimal.*

*S5: 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 $V$.*

The proof of Theorem 1, to be given in the Appendix, uses Lemma 1 which describes 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:*

*L1: $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}\ (P)$;*

*L2: $P(\mathcal{ABC}) = \frac{P(\mathcal{AC})P(\mathcal{BC})}{P(\mathcal{C})}$;*

*L3:* $\lambda_{\mathcal{D}}^{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(P) = 0, \quad \mathcal{D} \in \mathbb{D}(\mathcal{A}, \mathcal{B}, \mathcal{C})$*;*

*L4:* $P(\mathcal{ABC}) = t(\mathcal{AC})u(\mathcal{BC})$ *for some functions* $t$ *and* $u$*.*

The following examples illustrate applications and limitations of Theorem 1.

**Example 1** For the intersection of $A \perp\!\!\!\perp BC \mid DE$, $F \perp\!\!\!\perp BD \mid C$, $AF \perp\!\!\!\perp BE \mid DC$, $\mathbb{D}_2$, for example, is $\{FB, FD, FBD, FBC, FDC, FBDC\}$. Then, with $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3) = (ABCDE, BCDF, ABCDEF)$, condition (2) is satisfied, so the model is hierarchical marginal log-linear and smooth. A minimal specification of the model is:

$$\lambda_{\mathcal{E}}^{ABCDE} = 0, \ \mathcal{E} \in \mathbb{E}_1, \tag{5}$$
$$\lambda_{\mathcal{E}}^{ABCDE} = 0, \ \mathcal{E} \in \mathbb{E}_2, \tag{6}$$
$$\lambda_{\mathcal{E}}^{BCDF} = 0, \ \mathcal{E} \in \mathbb{E}_3, \tag{7}$$
$$\lambda_{\mathcal{E}}^{ABCDEF} = 0, \ \mathcal{E} \in \mathbb{E}_4, \tag{8}$$

with

$$\begin{aligned}\mathbb{E}_1 \ = \ \{&AB, AC, ABC, ABD, ACD, ABCD, ABE, ACE, ABCE, ABDE,\\ &ACDE, ABCDE\},\\ \mathbb{E}_2 = \{&AE, ADE\},\\ \mathbb{E}_3 = \{&BF, DF, BDF, BCF, CDF, BCDF\},\\ \mathbb{E}_4 = \{&EF, ABF, AEF, BEF, ABEF, DEF, ABDF,\\ &ADEF, BDEF, ABDEF, CEF, ABCF, ACEF, BCEF,\\ &ABCEF, CDEF, ABCDF, ACDEF, BCDEF, ABCDEF\}.\end{aligned}$$

By Lemma 1, (5) and (7) are equivalent to the first two conditional independencies defining the model. One might think that, in addition to (8), zero restrictions are needed for the log-linear parameters in $ABCDEF$ pertaining to the subsets $\mathcal{E} \in \mathbb{D}_3 \setminus \mathbb{E}_4$, as suggested by Lemma 1. But $\mathbb{D}_3 \setminus \mathbb{E}_4 \subset \mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3$, and Theorem 1 implies that these log-linear parameters need not be set to zero in $ABCDEF$. Thus, application of Theorem 1 is necessary to achieve minimal specification of the model. By calculating the number of linearly independent restrictions for each parameter in a minimal specification of the model, the number of degrees of freedom may be determined. □

**Example 2** The model defined as the intersection of $A \perp\!\!\!\perp B \mid D$, $A \perp\!\!\!\perp C \mid B$, $A \perp\!\!\!\perp D \mid C$ is not identified as a smooth model by Theorem 1, although from the inspection of the Jacobian of its parameterization we suspect that it is, in fact, smooth. □

## 3 Applications to graphical models

Graphical models associated with directed acyclic graphs (Lauritzen, 1996) are marginal log-linear models in the sense of Bergsma and Rudas (2002a), see Rudas, Bergsma, and Németh (2006). Here the Markov property is

$$V_i \perp\!\!\!\perp \mathrm{nd}(V_i) \mid \mathrm{pa}(V_i), \tag{9}$$

for every variable $V_i$, where $\mathrm{nd}(V_i)$ denotes the nondescendants and $\mathrm{pa}(V_i)$ denotes the parents of $V_i$. The marginal log-linear parameterization of such models given in Rudas et al. (2006) is based on a well-numbering of the variables (Lauritzen, Dawid, Larsen, and Leimer, 1990), such that (9) is equivalent to

$$V_i \perp\!\!\!\perp \mathrm{pre}(V_i) \setminus \mathrm{pa}(V_i) \mid \mathrm{pa}(V_i), \tag{10}$$

where $\mathrm{pre}(V_i)$ is the set of variables preceding $V_i$ in the well-numbering. The parameterization proposed by Rudas et al. (2006) is based on the marginals $\{V_i\} \cup \mathrm{pre}(V_i)$ for which (2) of Theorem 1 holds.

Statistical models associated with chain graphs have been considered, among others, by Lauritzen and Wermuth (1989), Frydenberg (1990), Cox and Wermuth (1996), Andersson et al. (2001), Richardson (2003), Wermuth and Cox (2004), Drton (2009).

For a component $\mathcal{K} \subseteq \mathcal{V}$ of a chain graph, $\mathrm{ND}(\mathcal{K})$ 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 neighbours of $\mathcal{X} \subseteq \mathcal{K}$, $\mathrm{nb}(\mathcal{X})$, is the set of nodes in $\mathcal{K}$ that are connected to a node in $\mathcal{X}$ and $\mathrm{pa}(\mathcal{X})$ is the set of nodes from which an arrow points to any node in $\mathcal{X}$.

Chain graph models are defined by combinations of some of the following properties.

P1: For all components $\mathcal{K}$, $\mathcal{K} \perp\!\!\!\perp \{\mathrm{ND}(\mathcal{K}) \setminus \mathrm{PA}(\mathcal{K})\} \mid \mathrm{PA}(\mathcal{K})$,
P2a: For all $\mathcal{K}$ and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X})\} \mid \{\mathrm{PA}(\mathcal{K}) \cup \mathrm{nb}(\mathcal{X})\}$,
P2b: For all $\mathcal{K}$ and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X})\} \mid \mathrm{PA}(\mathcal{K})$,

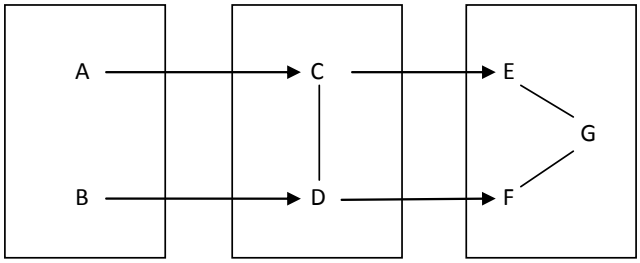


Figure 1: Chain graph whose Andersson–Madigan–Perlman interpretation is a smooth model (see Example 3)

P3a: For all $\mathcal{K}$ and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathrm{PA}(\mathcal{K}) \setminus \mathrm{pa}(\mathcal{X})\} \mid \{\mathrm{pa}(\mathcal{X}) \cup \mathrm{nb}(\mathcal{X})\}$,
P3b: For all $\mathcal{K}$ and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathrm{PA}(\mathcal{K}) \setminus \mathrm{pa}(\mathcal{X})\} \mid \mathrm{pa}(\mathcal{X})$.

The Type I Markov property (P1, P2a, P3a) is also called the Lauritzen–Wermuth–Frydenberg block-recursive Markov property, see Lauritzen and Wermuth (1989) and Frydenberg (1990), and the Type II Markov property (P1, P2a, P3b) is also called the Andersson–Madigan–Perlman block-recursive Markov property, see Andersson et al. (2001).

Smoothness of Type I models is implied by the results of Frydenberg (1990) and is also easily obtained applying Theorem 1.

The following example illustrates that Theorem 1 may be used to establish smoothness of chain graph models belonging to model classes which also contain nonsmooth models.

*Example 3.* The graph in Figure 1 with Type II interpretation is a smooth model and may be parameterized using the marginals $AB$, $ABC$, $ABD$, $CDE$, $CDF$, $CDG$, $CDEG$, $CDFG$, $CDEFG$, $ABCDEFG$. Type II models are not smooth in general, see Drton (2009), but in this case Theorem 1 implies smoothness immediately. □

Drton (2009) showed that Type IV models (P1, P2b, P3b) are smooth and gave a parameterization. Marchetti and Lupparelli (2008) illustrated through examples that these models are marginal log-linear. We now apply the general method in Theorem 1 to prove smoothness based on an interpretable parameterization, also implying the number of degrees of freedom associated with a Type IV model.

**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. Suppose the chain graph has components* $\mathcal{K}_1, \ldots, \mathcal{K}_T$, *that are well-numbered. The parameterization is based on the marginals*

$$\{\mathrm{PA}(\mathcal{K}_t) \cup \mathcal{X} : \mathcal{X} \subseteq \mathcal{K}_t\}^*, \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t, \ t = 1, \ldots, T, \tag{11}$$

*where* $\{\ \}^*$ *denotes a hierarchical ordering of the elements of the set. The parameters set to zero to define the model are those associated with the effects in*

$$\begin{aligned} &\{\mathbb{ID}(\mathcal{X}, \mathcal{K}_t \setminus \mathcal{X} \setminus \mathrm{nb}(\mathcal{X}), \mathrm{PA}(\mathcal{K}_t)) : \mathcal{X} \subseteq \mathcal{K}_t\} \cup \\ &\{\mathbb{ID}(\mathcal{X}, \mathrm{PA}(\mathcal{K}_t) \setminus \mathrm{pa}(\mathcal{X}), \mathrm{pa}(\mathcal{X})) : \mathcal{X} \subseteq \mathcal{K}_t\} \cup \\ &\qquad \mathbb{ID}(\mathcal{K}_t, (\mathcal{K}_t) \setminus \mathrm{PA}(\mathcal{K}_t), \mathrm{PA}(\mathcal{K}_t)), \end{aligned} \tag{12}$$

*for all components* $\mathcal{K}_t$, *where* $(\mathcal{K}_t)$ *is the set of components that precede* $\mathcal{K}_t$.

**Proof** For each component $\mathcal{K}_t$, the conditioning set in P2b is $\mathrm{PA}(\mathcal{K}_t)$ and in P3b it is $\mathrm{pa}(\mathcal{X}) \subseteq \mathrm{PA}(\mathcal{K}_t)$, thus for all conditional independencies implied by P2b or P3b, if written in the form of $\mathcal{A}_i \perp\!\!\!\perp \mathcal{B}_i \mid \mathcal{C}_i$, $\mathcal{C}_i \subseteq \mathrm{PA}(\mathcal{K}_t)$. Further, for these conditional independencies, $\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i = \mathrm{PA}(\mathcal{K}_t \cup \mathcal{X})$ for some $\mathcal{X} \subseteq \mathcal{K}_t$. Thus, for

$$\{\mathrm{PA}(\mathcal{K}_t) \cup \mathcal{X} : \mathcal{X} \subseteq \mathcal{K}_t\}^* \tag{13}$$

condition (2) holds and Theorem 1 applies. Therefore, a hierarchical marginal log-linear parameterization of the distributions with Properties P2b and P3b for any $\mathcal{K}_t$ is obtained. In addition, P1 has to be imposed.

The proof of equivalence between the local directed Markov property (9) and the local well-numbering Markov property (10) in Lauritzen et al. (1990) also applies to components of chain graphs, so it is also true that for a distribution on the chain graph, P1 holds if and only if the following P4 does.

P4: For all $\mathcal{K}_t$, $\mathcal{K}_t \perp\!\!\!\perp (\mathcal{K}_t) \setminus \mathrm{PA}(\mathcal{K}_t) \mid \mathrm{PA}(\mathcal{K}_t)$.

Because $(\mathcal{K}_t) = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{t-1}$, P4 may be parameterized using the marginal

$$\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_t. \tag{14}$$

Adding (14) after the marginals in (13) is hierarchical and the conditional independency in P4 is parameterized by setting to zero marginal log-linear parameters that are associated with effects appearing for the first time in (14), see Lemma 1.

Hence a hierarchical marginal log-linear parameterization of the Type IV block-recursive model may be obtained by using the marginals in (13) and (14), yielding (11). Theorem 1 implies that in the parameterization based

on the marginals in (11), the effects associated with (12) are zero and the remaining parameters parameterize the distributions in the model. □

As implied by *S3* of Theorem 1, the parameters not set to zero in (12) parameterize the model. These parameters are associated with the same effects as those found by Marchetti and Lupparelli (2008) to have nonzero parameters in the examples they investigated, although the marginals used for the parameterization are different.

## Acknowledgements

The first author is also a Recurrent Visiting Professor in the Central European University, Budapest, and the support received is acknowledged. The second author was supported by PASCAL travel grants. The authors thank Thomas Richardson and Michael Perlman for several discussions, and the reviewers for very helpful comments.

## Appendix

**Proof of Theorem 1** First we prove that *S1* is true. To see that $Q \in \mathcal{Q}_k$ implies (3), let $i \in \{1, \ldots, k\}$ and $\mathcal{E} \in \mathbb{D}_i$ be arbitrary. Because of (2),

$$[\mathcal{A}_i \cap \mathcal{M}(\mathcal{E})] \cup [\mathcal{B}_i \cap \mathcal{M}(\mathcal{E})] \cup \mathcal{C}_i$$

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

$$= [\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i] \cap \mathcal{M}(\mathcal{E}) = \mathcal{M}(\mathcal{E}).$$

Hence if $Q \in \mathcal{Q}_k$, then $\mathcal{A}_i \cap \mathcal{M}(\mathcal{E}) \perp\!\!\!\perp \mathcal{B}_i \cap \mathcal{M}(\mathcal{E}) \mid \mathcal{C}_i\,(Q)$, and because $\mathcal{E} \in \mathbb{D}(\mathcal{A}_i \cap \mathcal{M}(\mathcal{E}), \mathcal{B}_i \cap \mathcal{M}(\mathcal{E}), \mathcal{C}_i)$, *L3* implies (3).

To see that (3) implies $Q \in \mathcal{Q}_k$ define, for $j \leq m$, $\mathcal{I}_j = \{i \leq k \mid \mathcal{C}_i \subseteq \mathcal{M}_j \subseteq \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i\}$. The set $\mathcal{I}_j$ contains the indices in the specification of $\mathcal{Q}_k$ which may imply a conditional independence restriction for the marginal $\mathcal{M}_j$. Indeed, with $\mathcal{A}_{ij} = \mathcal{A}_i \cap \mathcal{M}_j$ and $\mathcal{B}_{ij} = \mathcal{B}_i \cap \mathcal{M}_j$, where either of these sets may be empty,

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

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

$$\mathcal{A}_{ij} \perp\!\!\!\perp \mathcal{B}_{ij} \mid \mathcal{C}_i\,(Q) \qquad i \in \mathcal{I}_j. \tag{15}$$

Since $\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i \in \mathbb{ID}_i$, by (2) 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 $j = j_i$ shows that if (15) holds for all $j \leq m$, then $Q \in \mathcal{Q}_k$, because $\mathcal{M}_m = \mathcal{V}$. To complete the proof of *S1*, it is therefore sufficient to show that (3) implies that (15) holds for all $j \leq m$.

Let $Q_j$ denote the restriction of $Q$ to $\mathcal{M}_j$. It is sufficient to show (15) for $Q_j$, which we do by applying two nested induction arguments. Because of (3) and *L3*, (15) is true for $Q_1$. Let the outer induction assumption be that (15) is true for $Q_j$ for all $j < l$, for some $l \leq m$ and now we prove (15) for all $j \leq l$.

Define a strictly positive probability distribution $P_l$ on $\mathcal{M}_l$, such that if $\mathcal{E} \subseteq \mathcal{M}_l \setminus (\cup_{j<l}\mathcal{M}_j)$, then $\lambda_{\mathcal{E}}^{\mathcal{M}_l}(P_l) = \lambda_{\mathcal{E}}^{\mathcal{M}_l}(Q_l)$ and that $\mathcal{A}_{il} \perp\!\!\!\perp \mathcal{B}_{il} \mid \mathcal{C}_i\,(P_l)$ holds for all $i \in \mathcal{I}_l$. These two requirements are compatible, because of (3) and *L3*.

Now characterize $Q_l$ by a mixed parameterization in the exponential family sense, see Barndorff-Nielsen (1978), Rudas (1998). Then $Q_l$ may be obtained from $P_l$ by replacing the marginal distributions of $P_l$ on all $\mathcal{M}_j \cap \mathcal{M}_l$, $j < l$ with $Q_j(\mathcal{M}_j \cap \mathcal{M}_l)$, without changing its log-linear parameters for the effects $\mathcal{E} \subseteq \mathcal{M}_l \setminus (\cup_{j<l}\mathcal{M}_j)$. To achieve this, the iterative proportional fitting procedure may be applied, starting with $P^0 = P_l$, for which (15) holds and it factorizes as in *L4* according to (15) for $j = l$. In every step, the IPFP adjusts one marginal. The adjustment in step $h$, for $h = 1, \ldots,$ is

$$P^h(\mathcal{M}_l) = P^{h-1}(\mathcal{M}_l)\frac{Q_j(\mathcal{M}_j \cap \mathcal{M}_l)}{P^{h-1}(\mathcal{M}_j \cap \mathcal{M}_l)}, \tag{16}$$

where $j = h(\mathrm{mod}\,(l-1))$. Let the inner induction assumption be that $P^{h-1}$ factorizes as in *L4* according to (15) for $j = l$, which is true for $h = 1$. Then $P^h$ factorizes as well, because all its factors in (16) do so. Indeed, if $i \in \mathcal{I}_j$, the outer induction assumption is that $Q_j$ factorizes and so does $Q_j(\mathcal{M}_j \cap \mathcal{M}_l)$. If $i \in \mathcal{I}_l \setminus \mathcal{I}_j$, then either $\mathcal{M}_j \cap \mathcal{M}_l \notin \mathbb{ID}_i$ and the factorization is trivial, or $\mathcal{C}_i \nsubseteq \mathcal{M}_j$ (otherwise $i \in \mathcal{I}_j$ would follow), and although $\mathcal{M}_j \supseteq \mathcal{M}_j \cap \mathcal{M}_l$, $\mathcal{M}_j$ cannot be $\mathcal{M}(\mathcal{M}_j \cap \mathcal{M}_l)$ because of (2), thus $\mathcal{M}(\mathcal{M}_j \cap \mathcal{M}_l) = \mathcal{M}_{j'}$ for some $j' < j$. $Q_{j'}$ factorizes as required by the outer induction assumption, and $\mathcal{M}_j \cap \mathcal{M}_l \subseteq \mathcal{M}_{j'}$, so $Q_j(\mathcal{M}_j \cap \mathcal{M}_l) = Q_{j'}(\mathcal{M}_j \cap \mathcal{M}_l)$ also factorizes.

As implied by Csiszár (1975), the procedure converges to $Q_l$ and by positivity also the limit factorizes as in *L4* according to (15) for $j = l$, which completes the inner induction. Because of the outer induction assumption, for all $j < l$, (15) already holds for $Q_j$, thus the outer induction step and with it the proof of *S1* is completed.

To see the rest of the Theorem, note that, since $\mathcal{M}_1, \ldots, \mathcal{M}_m$ is hierarchical and complete, Theorem 2 of Bergsma and Rudas (2002a) can be applied with, using the notation of that paper, $\mathcal{P}$ and $\tilde{\lambda}_{\mathcal{P}}$ defined as $\mathcal{P} = \{(\mathcal{E}, \mathcal{M}(\mathcal{E})) \mid \mathcal{E} \subseteq \mathcal{V}\}$ and $\tilde{\lambda}_{\mathcal{P}} = \{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} \mid \mathcal{E} \subseteq \mathcal{V}\}$, implying that the latter is a smooth parameterization of the distributions on the contingency table, implying *S2, S3, S4*. Theorem 5 in Bergsma and Rudas (2002a) can now be applied to obtain *S5*. □

## 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., & 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.

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.

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

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

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.

Marchetti, G. L., & Lupparelli, M. (2008). Parameterization and fitting of a class of discrete graphical models. In P. Brito (Ed.), *Compstat 2008, Proceedings in Computational Statistics* (p. 117-128). Heidelberg: Physica Verlag.

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. (2005). *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.