Bernoulli **17**(3), 2011, 827–844
DOI: 10.3150/10-BEJ300

Chain graph models of multivariate regression type for categorical data

GIOVANNI M. MARCHETTI[1] and MONIA LUPPARELLI[2]

[1]*Dipartimento di Statistica "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Firenze, Italy. E-mail: giovanni.marchetti@ds.unifi.it*
[2]*Dipartimento di Scienze Statistiche "P. Fortunati", University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy. E-mail: monia.lupparelli@unibo.it*

We discuss a class of chain graph models for categorical variables defined by what we call a multivariate regression chain graph Markov property. First, the set of local independencies of these models is shown to be Markov equivalent to those of a chain graph model recently defined in the literature. Next we provide a parametrization based on a sequence of generalized linear models with a multivariate logistic link function that captures all independence constraints in any chain graph model of this kind.

Keywords: block-recursive Markov property; discrete chain graph models of type IV; graphical Markov models; marginal log-linear models; multivariate logistic regression models

1. Introduction

Discrete graphical Markov models are models for discrete distributions representable by graphs, associating nodes with the variables and using rules that translate properties of the graph into conditional independence statements between variables. There are several classes of graphical models; see [24] for a review. In this paper we focus on the class of multivariate regression chain graphs and we discuss their definition and parametrization for discrete variables.

Multivariate regression chain graphs generalize directed acyclic graphs, which model recursive sequences of univariate responses, by allowing multiple responses. As in all chain graph models the variables can be arranged in a sequence of blocks, called chain components, ordered on the basis of subject-matter considerations, and the variables within a block are considered to be on an equal standing as responses. The edges are undirected within the chain components, drawn as dashed lines [6] or as bi-directed arrows [21], and directed between components, all pointing in the same direction, that is, with no chance of semi-directed cycles. One special feature of multivariate regression chain graphs is that the responses are potentially depending on all the variables in all previous groups, but not on the other responses. Chain graphs with this interpretation were proposed first by Cox and Wermuth in [5], with several examples in [6], Chapter 5.

In the special case of a single group of responses with no explanatory variables, multivariate regression chain graphs reduce to covariance graphs, that is, to undirected graphs representing marginal independencies with the basic rule that if its subgraph is disconnected, that is, composed by completely separated sets of nodes, then the associated variables are jointly independent; see [11] and [18]. In the general case, the interpretation of the undirected graphs within a chain component is that of a covariance graph, but conditional on all variables in preceding components.

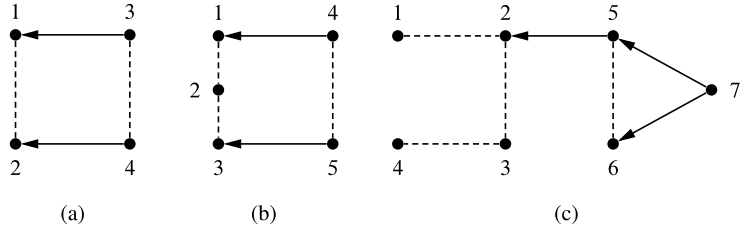

Figure 1. Three chain graphs with chain components (a) $\mathcal{T} = \{\{1, 2\}, \{3, 4\}\}$; (b) $\mathcal{T} = \{\{1, 2, 3\}, \{4, 5\}\}$; (c) $\mathcal{T} = \{\{1, 2, 3, 4\}, \{5, 6\}, \{7\}\}$. Dashed lines only occur within chain components.

For example, the missing edge $(1, 3)$ in the graph of Figure 1(b) is interpreted as the independence statement $X_1 \perp\!\!\!\perp X_3 | X_4, X_5$, compactly written in terms of nodes as $1 \perp\!\!\!\perp 3|4, 5$.

The interpretation of the directed edges is that of multivariate regression models, with a missing edge denoting a conditional independence of the response on a variable given all the remaining potential explanatory variables. Thus, in the chain graph of Figure 1(a) the missing arrow $(1, 4)$ indicates the independence statement $1 \perp\!\!\!\perp 4|3$. The interpretation differs from that of classical chain graphs ([12,17]; LWF for short) where the missing edges mean conditional independencies given all the remaining variables, including the other responses within the same block. However, in studies involving longitudinal data, such as the prospective study of child development discussed in [4], where there are blocks of joint responses recorded at ages of three months, two years and four years, an analysis conditioning exclusively on the previous developmental states is typically appropriate.

Recently, [8] distinguished four types of chain graphs comprising the classical and the alternative [1] chain graph models, called type I and II, respectively. In this paper we give a formal definition of multivariate regression chain graph models and we prove that they are equivalent to the chain graph models of type IV, in Drton's classification [8]. Moreover, we provide a parametrization based on recursive multivariate logistic regression models. These models, introduced in [20], Section 6.5.4, and [13] can be used to define all the independence constraints. The models can be defined by an intuitive rule, see Theorem 2, based on the structure of the chain graph, that can be translated into a sequence of explicit regression models. One consequence of the given results is that any discrete multivariate regression chain graph model is a curved exponential family, a result obtained in [8] with a different proof.

2. Multivariate regression chain graphs

The basic definitions and notation used in this paper closely follow [8], and they are briefly recalled below. A chain graph $G = (V, E)$ is a graph with finite node set $V = \{1, \ldots, d\}$ and an edge set E that may contain either directed edges or undirected edges. The graph has no semi-directed cycle, that is, no path from a node to itself with at least one directed edge such that all directed edges have the same orientation. The node set V of a chain graph can be partitioned into disjoint subsets $T \in \mathcal{T}$ called *chain components*, such that all edges in each subgraph G_T are undirected and the edges between different subsets $T_1 \neq T_2$ are directed, pointing in the same

direction. For chain graphs with the multivariate regression interpretation, the subgraphs G_T within each chain component have undirected dashed (- - -) or bi-directed ($\longleftrightarrow$) edges. The former convention is adopted in this paper. Thus, the chain graph of Figure 1(c) has three chain components, while the previous ones have two components.

Given a subset $A \subseteq T$ of nodes within a chain component, the subgraph G_A is said to be *disconnected* if there exist two nodes in A such that no path in G_A has those nodes as endpoints. In this case, A can be partitioned uniquely into a set of $r > 1$ connected components $A_1, \ldots, A_r$. Otherwise, the subgraph G_A is *connected*. For example, in chain graph (c) of Figure 1, the subgraph G_A with $A = \{1, 2, 4\}$ is disconnected with two connected components $A_1 = \{1, 2\}$ and $A_2 = \{4\}$. On the other hand, the subgraph G_A with $A = \{1, 2, 3\}$ is connected. In the remainder of the paper, we shall say for short that a subset A of nodes in a component is connected (respectively, disconnected) if the subgraph G_A is connected (respectively, disconnected).

Any chain graph yields a directed acyclic graph D of its chain components having $\mathcal{T}$ as a node set and an edge $T_1 \longrightarrow T_2$ whenever there exists in the chain graph G at least one edge $v \longrightarrow w$ connecting a node v in T_1 with a node w in T_2. In this directed graph, we may define for each T the set $\mathrm{pa}_D(T)$ as the union of all the chain components that are parents of T in the directed graph D. This concept is distinct from the usual notion of the *parents* $\mathrm{pa}_G(A)$ of a set of nodes A in the chain graph, that is, the set of all the nodes w outside A such that $w \longrightarrow v$ with $v \in A$. For instance, in the graph of Figure 2(a), for $T = \{1, 2\}$, the set of parent components is $\mathrm{pa}_D(T) = \{3, 4, 5, 6\}$, whereas the set of parents of T is $\mathrm{pa}_G(T) = \{3, 6\}$.

In this paper we start the analysis from a given chain graph $G = (V, E)$ with an associated collection $\mathcal{T}$ of chain components. However, in applied work, where variables are linked to nodes by the correspondence X_v for $v \in V$, usually a set of chain components is assumed known from previous studies of substantive theories or from the temporal ordering of the variables. For variables within such chain components no direction of influence is specified and they are considered as joint responses, that is, to be on equal standing. The relations between variables in different chain components are directional and are typically based on a preliminary distinction of responses, intermediate responses and purely explanatory factors. Often, a full ordering of the components is assumed based on time order or on a subject matter working hypothesis; see [6].

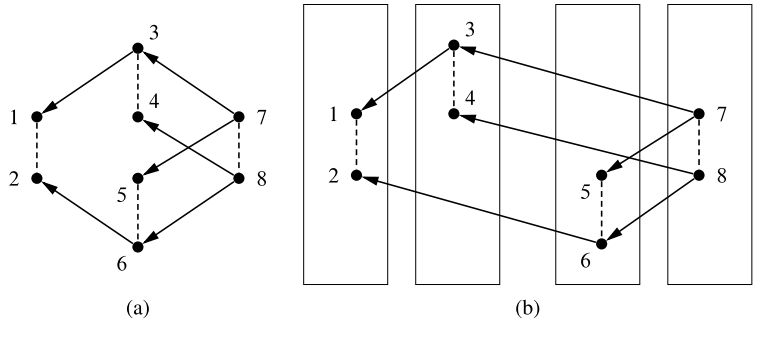

Figure 2. (a) A chain graph and (b) one possible consistent ordering of the four chain components: $\{1, 2\} \prec \{3, 4\} \prec \{5, 6\} \prec \{7, 8\}$. In (b) the set of predecessors of $T = \{1, 2\}$ is $\mathrm{pre}(T) = \{3, 4, 5, 6, 7, 8\}$, while the set of parent components of T is $\mathrm{pa}_D(T) = \{3, 4, 5, 6\}$.

Given a chain graph G with chain components $(T \mid T \in \mathcal{T})$, we can always define a strict total order $\prec$ of the chain components that is *consistent* with the partial order induced by the chain graph, such that if $T \prec T'$ then $T \notin \mathrm{pa}_D(T')$. For instance, in the chain graph of Figure 2(a) there are four chain components ordered in graph (b) as $\{1,2\} \prec \{3,4\} \prec \{5,6\} \prec \{7,8\}$. Note that the chosen total order of the chain components is in general not unique and that another consistent ordering could be $\{1,2\} \prec \{5,6\} \prec \{3,4\} \prec \{7,8\}$.

In the remainder of the paper we shall assume that a consistent ordering $\prec$ of the chain components is given. Then, for each T, the set of all components preceding T is known and we may define the cumulative set $\mathrm{pre}(T) = \bigcup_{T' \prec T} T'$ of nodes contained in the predecessors of component T that we sometimes also call the past of T. The set $\mathrm{pre}(T)$ captures the notion of all the potential explanatory variables of the response variables within T. By definition, as the full ordering of the components is consistent with G, the set of predecessors $\mathrm{pre}(T)$ of each chain component T always includes the parent components $\mathrm{pa}_D(T)$.

The following definition explains the meaning of the multivariate regression interpretation of a chain graph.

Definition 1. *Let G be a chain graph with chain components $(T \mid T \in \mathcal{T})$ and let* $\mathrm{pre}(T)$ *define an ordering of the chain components consistent with the graph. A joint distribution P of the random vector $\mathbf{X}$ obeys the* (*global*) *multivariate regression Markov property for G if it satisfies the following independencies. For all $T \in \mathcal{T}$ and for all $A \subseteq T$:*

(MR1) *if A is connected*: $A \perp\!\!\!\perp [\mathrm{pre}(T) \setminus \mathrm{pa}_G(A)] \mid \mathrm{pa}_G(A)$.
(MR2) *if A is disconnected with connected components* $A_1, \ldots, A_r$: $A_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp A_r \mid \mathrm{pre}(T)$.

Assuming that the distribution P has a density p with respect to a product measure, the definition can be stated by the following two equivalent conditions:

$$p_{A|\,\mathrm{pre}(T)} = p_{A|\,\mathrm{pa}_G(A)} \tag{1a}$$

for all T and for all connected subset $A \subseteq T$.

$$p_{A|\,\mathrm{pre}(T)} = \prod_j p_{A_j|\,\mathrm{pre}(T)} \tag{1b}$$

for all T and for all disconnected subset $A \subset T$ with connected components A_j, $j = 1, \ldots, r$.

In other words, for any connected subset A of responses in a component T, its conditional distribution given the variables in the past depends only on the parents of A. On the other hand, if A is disconnected (i.e., the subgraph G_A is disconnected) the variables in its connected components $A_1, \ldots, A_r$, are jointly independent given the variables in the past.

Remark 1. Definition 1 gives a local Markov property that always implies the following pairwise Markov property: For every uncoupled pair of nodes i, k,

$$i \perp\!\!\!\perp k|\,\mathrm{pre}(T), \qquad \text{if } i,k \in T; \qquad i \perp\!\!\!\perp k|\,\mathrm{pre}(T) \setminus \{k\}, \qquad \text{if } i \in T, k \in \mathrm{pre}(T). \tag{2}$$

In particular, two pairwise independencies $i \perp\!\!\!\perp k | \operatorname{pre}(T)$ and $i \perp\!\!\!\perp \ell | \operatorname{pre}(T)$ can occur only in combination with the joint independence $i \perp\!\!\!\perp k, \ell | \operatorname{pre}(T)$. This means that in the associated model the composition property is always satisfied; see [22]. Thus, even though we concentrate in this paper on the family of multinomial distributions that does not satisfy the composition property, the models in which (MR1) and (MR2) hold have this property.

Remark 2. One immediate consequence of Definition 1 is that if the probability density $p(\mathbf{x})$ is strictly positive, then it factorizes according to the directed acyclic graph of the chain components:

$$p(\mathbf{x}) = \prod_{T \in \mathcal{T}} p\big(\mathbf{x}_T | \mathbf{x}_{\operatorname{pa}_D(T)}\big). \tag{3}$$

This factorization property is shared by all types of chain graphs; see [24] and [8].

Recently, [8] discussed four different block-recursive Markov properties for chain graphs, of which we discuss here those with the *Markov property of type IV*. To state it, we need two further concepts from graph theory. Given a chain graph G, the set $\operatorname{Nb}_G(A)$ is the union of A itself and the set of nodes w that are *neighbours* of A, that is, coupled by an undirected edge to some node v in A. Moreover, the set of *non-descendants* $\operatorname{nd}_D(T)$ of a chain component T, is the union of all components T' such that there is no directed path from T to T' in the directed graph of chain components D.

Definition 2 (Chain graph Markov property of type IV [8]). *Let G be a chain graph with chain components $(T \mid T \in \mathcal{T})$ and directed acyclic graph D of components. The joint probability distribution of $\mathbf{X}$ obeys the block-recursive Markov property of type IV if it satisfies the following independencies*:

(IV0) $A \perp\!\!\!\perp [\operatorname{nd}_D(T) \setminus \operatorname{pa}_D(T)] \mid \operatorname{pa}_D(T)$ *for all* $T \in \mathcal{T}$;
(IV1) $A \perp\!\!\!\perp [\operatorname{pa}_D(T) \setminus \operatorname{pa}_G(A)] \mid \operatorname{pa}_G(A)$ *for all* $T \in \mathcal{T}$ *for all* $A \subseteq T$;
(IV2) $A \perp\!\!\!\perp [T \setminus \operatorname{Nb}_G(A)] \mid \operatorname{pa}_D(T)$ *for all* $T \in \mathcal{T}$ *for all connected subsets* $A \subseteq T$.

Then we have the following result, proved in the Appendix.

Theorem 1. *Given a chain graph G, the multivariate regression Markov property is equivalent to the block-recursive Markov property of type IV.*

This result shows that the block-recursive property of a chain graph of type IV is in fact simplified by Definition 1. On the other hand, Definition 1 depends only apparently on the chosen full ordering of the chain components, because the equivalent Definition 2 depends only on the underlying chain graph G.

Example 1. The independencies implied by the multivariate regression chain graph Markov property are illustrated below for each of the graphs of Figure 1.

Graph (a) represents the independencies of the seemingly unrelated regression model; see [5] and [10]. For $T = \{1, 2\}$ and $\operatorname{pre}(T) = \{3, 4\}$ we have the independencies $1 \perp\!\!\!\perp 4|3$ and $2 \perp\!\!\!\perp 3|4$.

Note that for the connected set $A = \{1, 2\}$ the condition (MR1) implies the trivial statement $A \perp\!\!\!\perp \varnothing | \operatorname{pre}(T)$.

In graph (b) one has $T = \{1, 2, 3\}$ and $\operatorname{pre}(T) = \{4, 5\}$. Thus, for each connected subset $A \subseteq T$, by (MR1), we have the non-trivial statements

$$1 \perp\!\!\!\perp 5|4; \qquad 2 \perp\!\!\!\perp 4, 5; \qquad 3 \perp\!\!\!\perp 4|5; \qquad 1, 2 \perp\!\!\!\perp 5|4; \qquad 2, 3 \perp\!\!\!\perp 4|5.$$

Then, for the remaining disconnected set $A = \{1, 3\}$ we obtain by (MR2) the independence $1 \perp\!\!\!\perp 3|4, 5$.

In graph (c), considering the conditional distribution $p_{T|\operatorname{pre}(T)}$ for $T = \{1, 2, 3, 4\}$ and $\operatorname{pre}(T) = \{5, 6, 7\}$, we can define independencies for each of the eight connected subsets of T. For instance, we have

$$1 \perp\!\!\!\perp 5, 6, 7; \qquad 1, 2 \perp\!\!\!\perp 6, 7|5; \qquad 1, 2, 3, 4 \perp\!\!\!\perp 7|5, 6.$$

The last independence is equivalent to the factorization $p = p_{1234|56} \cdot p_{56|7} \cdot p_7$ of the joint probability distribution according to the directed acyclic graph of the chain components. The remaining five disconnected subsets of T imply the conditional independencies $1, 2 \perp\!\!\!\perp 4|5, 6, 7$ and $1 \perp\!\!\!\perp 3, 4|5, 6, 7$. Notice that when in a component there are two uncoupled nodes, then there is a conditional independence given simply the common parents of the two nodes. For example, in graph (c), we have not only $1 \perp\!\!\!\perp 3|5, 6$ but also $1 \perp\!\!\!\perp 3|5$.

Remark 3. When each component T induces a complete subgraph G_T and, for all subsets A in T, the set of parents of A, $\operatorname{pa}_G(A)$, coincides with the set of the parent components of T, $\operatorname{pa}_D(T)$, then the only conditional independence implied by the multivariate regression Markov property is

$$A \perp\!\!\!\perp [\operatorname{pre}(T) \setminus \operatorname{pa}_D(T)] | \operatorname{pa}_D(T) \qquad \text{for all } A \subseteq T,\ T \in \mathcal{T}.$$

This condition is in turn equivalent just to the factorization (3) of the joint probability distribution.

Remark 4. In Definition 1, (MR2) is equivalent to imposing that for all T the conditional distribution $p_{T|\operatorname{pre}(T)}$ satisfies the independencies of a covariance graph model with respect to the subgraph G_T.

In [18], Proposition 3, it is shown that a covariance graph model is defined by constraining to zero, in the multivariate logistic parametrization, the parameters corresponding to all disconnected subsets of the graph. In the following subsection we extend this approach to the multivariate regression chain graph models.

3. Recursive multivariate logistic regression models

3.1. Notation

Let $\mathbf{X} = (X_v \mid v \in V)$ be a discrete random vector, where each variable X_v has a finite number r_v of levels. Thus $\mathbf{X}$ takes values in the set $\mathcal{I} = \prod_{v \in V} \{1, \ldots, r_v\}$ whose elements are the cells of the

joint contingency table, denoted by $\mathbf{i} = (i_1, \ldots, i_d)$. The first level of each variable is considered a reference level and we consider also the set $\mathcal{I}^\star = \prod_{v\in V}\{2, \ldots, r_v\}$ of cells having all indices different from the first. The elements of $\mathcal{I}^\star$ are denoted by $\mathbf{i}^\star$.

The joint probability distribution of $\mathbf{X}$ is defined by the mass function

$$p(\mathbf{i}) = P(X_v = i_v, v = 1, \ldots, d) \qquad \text{for all } \mathbf{i} \in \mathcal{I},$$

or equivalently by the probability vector $\mathbf{p} = (p(\mathbf{i}), \mathbf{i} \in \mathcal{I})$. With three variables we shall use often p_{ijk} instead of $p(i_1, i_2, i_3)$.

Given two disjoint subsets A and B of V, the marginal probability distribution of $\mathbf{X}_B$ is $p(\mathbf{i}_B) = \sum_{\mathbf{j}_B=\mathbf{i}_B} p(\mathbf{j})$ where $\mathbf{i}_B$ is a subvector of $\mathbf{i}$ belonging to the marginal contingency table $\mathcal{I}_B = \prod_{v\in B}\{1, \ldots, r_v\}$. The conditional probability distributions are defined as usual and denoted by $p(\mathbf{i}_A|\mathbf{i}_B)$, for $\mathbf{i}_A \in \mathcal{I}_A$ and $\mathbf{i}_B \in \mathcal{I}_B$ or, compactly, by $p_{A|B}$. When appropriate, we define the set $\mathcal{I}_B^\star = \prod_{v\in B}\{2, \ldots, r_v\}$.

A discrete multivariate regression chain graph model $\mathbf{P}_{\mathrm{MR}}(G)$ associated with the chain graph $G = (V, E)$ is the set of strictly positive joint probability distributions $p(\mathbf{i})$ for $i \in \mathcal{I}$ that obeys the multivariate regression Markov property. By Theorem 1 this class coincides with the set $\mathbf{P}_{\mathrm{IV}}(G)$ of discrete chain graph models of type IV.

In the next subsection we define an appropriate parametrization for each component of the standard factorization

$$p(\mathbf{i}) = \prod_{T\in\mathcal{T}} p\big(\mathbf{i}_T|\mathbf{i}_{\mathrm{pre}(T)}\big) \tag{4}$$

of the joint probability distribution. Actually we define a saturated linear model for a suitable transformation of the parameters of each conditional probability distribution $p(\mathbf{i}_T|\mathbf{i}_{\mathrm{pre}(T)})$.

3.2. Multivariate logistic contrasts

The suggested link function is the multivariate logistic transformation; see [20], page 219, and [13]. This link transforms the joint probability vector of the responses into a vector of logistic contrasts defined for all the marginal distributions. The contrasts of interest are all sets of univariate, bivariate and higher order contrasts. In general, a *multivariate logistic contrast* for a marginal table $\mathbf{p}_A$ is defined by the function

$$\eta^{(A)}(\mathbf{i}_A^\star) = \sum_{s\subseteq A}(-1)^{|A\setminus s|}\log p(\mathbf{i}_s^\star, \mathbf{1}_{A\setminus s}) \qquad \text{for } \mathbf{i}^\star \in \mathcal{I}_A^\star, \tag{5}$$

where the notation $|A \setminus s|$ denotes the cardinality of set $A \setminus s$. The contrasts for a margin A are denoted by $\boldsymbol{\eta}^{(A)}$ and the full vector of the contrasts for all non-empty margins $A \subseteq V$ are denoted by $\boldsymbol{\eta}$. The following example illustrates the transformation for two responses.

Example 2. Let p_{ij}, for $i = 1, 2$, $j = 1, 2, 3$ be a joint bivariate distribution for two discrete variables X_1 and X_2. Then the multivariate logistic transform changes the vector $\mathbf{p}$ of probabilities,

belonging to the 5-dimensional simplex, into the 5×1 vector

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}^{(1)} \\ \boldsymbol{\eta}^{(2)} \\ \boldsymbol{\eta}^{(12)} \end{pmatrix}, \qquad \text{where } \boldsymbol{\eta}^{(1)} = \log \frac{p_{2+}}{p_{1+}}, \boldsymbol{\eta}^{(2)} = \begin{pmatrix} \log \dfrac{p_{+2}}{p_{+1}} \\ \log \dfrac{p_{+3}}{p_{+1}} \end{pmatrix}, \boldsymbol{\eta}^{(12)} = \begin{pmatrix} \log \dfrac{p_{11}p_{22}}{p_{21}p_{12}} \\ \log \dfrac{p_{11}p_{23}}{p_{21}p_{13}} \end{pmatrix},$$

where the $+$ suffix indicates summing over the corresponding index. Thus, the parameters $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$ are marginal baseline logits for the variables X_1 and X_2, while $\boldsymbol{\eta}^{(12)}$ is a vector of log odds ratios. The definition used in this paper uses baseline coding, that is, the contrasts are defined with respect to a reference level, by convention the first. Therefore the dimension of the vectors $\boldsymbol{\eta}^{(1)}$, $\boldsymbol{\eta}^{(2)}$ and $\boldsymbol{\eta}^{(12)}$ are the dimensions of the sets $\mathcal{I}_1^\star$, $\mathcal{I}_2^\star$ and $\mathcal{I}_{12}^\star$. Other coding schemes can be adopted, as discussed, for instance, in [23] and [2].

Remark 5. This transformation for multivariate binary variables is discussed in [13], where it is shown that the function from $\mathbf{p}$ to $\boldsymbol{\eta}$ is a smooth (C^∞) one-to-one function having a smooth inverse, that is, it is a diffeomorphism; see also [3]. For general discrete variables, see [18]. The parameters are not variation-independent, that is, they do not belong to a hyper-rectangle. However, they satisfy the *upward compatibility property*, that is, they have the same meaning across different marginal distributions; see [13] and [18], Proposition 4. Often the multivariate logistic link is written as

$$\boldsymbol{\eta} = \mathbf{C} \log(\mathbf{M}\mathbf{p}), \tag{6}$$

where $\mathbf{C}$ and $\mathbf{M}$ are suitable Kronecker products of contrast and marginalization matrices, respectively. For the explicit construction of these matrices, see [2].

3.3. Saturated model

We specify the dependence of the responses in each component T on the variables in the past by defining a saturated multivariate logistic model for the conditional probability distribution $p_{T|\,\mathrm{pre}(T)}$. The full saturated model for the joint probability p then follows from the factorization (4).

For each covariate class $\mathbf{i}_{\mathrm{pre}(T)} \in \mathcal{I}_{\mathrm{pre}(T)}$, let $\mathbf{p}(\mathbf{i}_{\mathrm{pre}(T)})$ be the vector with strictly positive components $p(\mathbf{i}_T|\mathbf{i}_{\mathrm{pre}(T)}) > 0$ for $\mathbf{i}_T \in \mathcal{I}_T$. Then consider the associated conditional multivariate logistic parameters $\boldsymbol{\eta}(\mathbf{i}_{\mathrm{pre}(T)})$ defined using the link function (6). Notice that this vector is composed of contrasts $\boldsymbol{\eta}^{(A)}(\mathbf{i}_{\mathrm{pre}(T)})$ for all non-empty subsets A of T. Then we express the dependence of each of them on the variables in the preceding components by a complete factorial model

$$\boldsymbol{\eta}^{(A)}\big(\mathbf{i}_{\mathrm{pre}(T)}\big) = \sum_{b \subseteq \mathrm{pre}(T)} \boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b). \tag{7}$$

Here the vectors $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b)$ have dimensions of the sets $\mathcal{I}_A^\star$, and are defined according to the baseline coding, and thus vanish when at least one component of $\mathbf{i}_b$ takes on the first level. Again, here different codings may be used if appropriate. Often it is useful to express (7) in matrix form

$$\boldsymbol{\eta}^{(A)} = \mathbf{Z}^{(A)} \boldsymbol{\beta}^{(A)}, \tag{8}$$

where $\boldsymbol{\eta}^{(A)}$ is the column vector obtained by stacking all vectors $\boldsymbol{\eta}^{(A)}(\mathbf{i}_{\mathrm{pre}(T)})$ for $\mathbf{i}_{\mathrm{pre}(T)} \in \mathcal{I}_{\mathrm{pre}(T)}$, $\mathbf{Z}^{(A)}$ is a full-rank design matrix and $\boldsymbol{\beta}^{(A)}$ is a parameter vector.

Example 3. Suppose that in Example 2 the responses X_1 and X_2 depend on two binary explanatory variables X_3 and X_4, with levels indexed by k and ℓ, respectively. Then the saturated model is

$$\boldsymbol{\eta}^{(A)}(k,\ell) = \boldsymbol{\beta}_{\varnothing}^{(A)} + \boldsymbol{\beta}_{3}^{(A)}(k) + \boldsymbol{\beta}_{4}^{(A)}(\ell) + \boldsymbol{\beta}_{34}^{(A)}(k,\ell), \qquad k,\ell = 1,2,$$

for $A = \{1\}, \{2\}, \{12\}$. The explicit form of the matrix $\mathbf{Z}^{(A)}$ in equation (8) is, using the Kronecker product $\otimes$ operator,

$$\mathbf{Z}^{(A)} = \mathbf{I} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

that is, a matrix of a complete factorial design matrix, where $\mathbf{I}$ is an identity matrix of an order equal to the common dimension of each $\boldsymbol{\eta}^{(A)}(k,\ell)$. Following [20], page 222, we shall denote the model, for the sake of brevity, by a multivariate model formula

$$X_1 : X_3 * X_4; \qquad X_2 : X_3 * X_4; \qquad X_{12} : X_3 * X_4,$$

where $X_3 * X_4 = X_3 + X_4 + X_3 \cdot X_4$ is the factorial expansion in Wilkinson and Rogers' symbolic notation [25].

When we need to express the overall 1–1 smooth transformation of the conditional probability vectors $\mathbf{p}(\mathbf{i}_{\mathrm{pre}(T)})$, denoted collectively by $\mathbf{p}_T$, into the logistic and regression parameters we introduce the vectors $\boldsymbol{\eta}_T$ and $\boldsymbol{\beta}_T$ obtained by concatenating the parameters $\boldsymbol{\eta}^{(A)}$ and $\boldsymbol{\beta}^{(A)}$, respectively, for all non-empty subsets A of T, writing

$$\boldsymbol{\eta}_T = \mathbf{Z}_T \boldsymbol{\beta}_T, \tag{9}$$

where $\mathbf{Z}_T = \mathrm{diag}(\mathbf{Z}^{(A)})$ is a full rank block-diagonal matrix of the saturated model, and

$$\mathbf{C}_T \log(\mathbf{M}_T \mathbf{p}_T) = \boldsymbol{\eta}_T, \tag{10}$$

where $\mathbf{C}_T$ and $\mathbf{M}_T$ are suitable overall contrast and marginalization matrices.

4. Discrete multivariate regression chain graph models

4.1. Linear constraints

A multivariate regression chain graph model is specified by zero constraints on the parameters $\boldsymbol{\beta}_T$ of the saturated model (9). We give first an example and then we state the general result.

Example 4. Continuing the previous example for the chain graph G of Figure 1(a), we shall require that X_1 depends only on X_3 and X_2 depends only on X_4. Therefore, we specify the model

$$\begin{aligned}\boldsymbol{\eta}^{(1)}(k,\ell) &= \boldsymbol{\beta}_{\varnothing}^{(1)} + \boldsymbol{\beta}_{3}^{(1)}(k),\\ \boldsymbol{\eta}^{(2)}(k,\ell) &= \boldsymbol{\beta}_{\varnothing}^{(2)} + \boldsymbol{\beta}_{4}^{(2)}(\ell),\\ \boldsymbol{\eta}^{(12)}(k,\ell) &= \boldsymbol{\beta}_{\varnothing}^{(12)} + \boldsymbol{\beta}_{3}^{(12)}(k) + \boldsymbol{\beta}_{4}^{(12)}(\ell) + \boldsymbol{\beta}_{34}^{(12)}(k,\ell)\end{aligned}$$

with a corresponding multivariate model formula

$$X_1 : X_3, \qquad X_2 : X_4, \qquad X_{12} : X_3 * X_4.$$

The reduced model satisfies the two independencies $1 \perp\!\!\!\perp 4|3$ and $2 \perp\!\!\!\perp 3|4$ because the first two equations are equivalent to $p_{1|34} = p_{1|3}$ and $p_{2|34} = p_{1|4}$, respectively. The log odds-ratio between X_1 and X_2, on the other hand, depends in general on all the combinations (k, ℓ) of levels of the two explanatory variables.

The following theorem, proved in the Appendix, states a general rule to parametrize any discrete chain graph model of the multivariate regression type.

Theorem 2. *Let G be a chain graph and let* $\mathrm{pre}(T)$ *be a consistent ordering of the chain components $T \in \mathcal{T}$. A joint distribution of the discrete random vector* $\mathbf{X}$ *belongs to* $\mathbf{P}_{\mathrm{MR}}(G)$ *if and only if, in the multivariate logistic model* (7)*, the parameters* $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b) = \mathbf{0}$, $\mathbf{i}_b \in \mathcal{I}_b$, *whenever*

$$A \text{ is connected and } b \subseteq \mathrm{pre}(T) \setminus \mathrm{pa}_G(A), \tag{11a}$$

$$A \text{ is disconnected and } b \subseteq \mathrm{pre}(T) \tag{11b}$$

for all $A \subseteq T$ and for all $T \in \mathcal{T}$.

Notice that equations (11a) and (11b) correspond to conditions (1a) and (1b), respectively, of Definition 1. Thus the multivariate regression chain graph model turns out to be $\boldsymbol{\eta}^{(A)}(\mathbf{i}_{\mathrm{pre}(T)}) = \sum_{b \subseteq \mathrm{pa}_G(A)} \boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b)$ if A is connected and $\mathbf{0}$ if A is disconnected. In matrix form we have a linear predictor

$$\boldsymbol{\eta}_T = \mathbf{Z}_r \boldsymbol{\beta}_r, \tag{12}$$

where $\mathbf{Z}_r$ is the matrix of the reduced model obtained by removing selected columns of $\mathbf{Z}_T$, and $\boldsymbol{\beta}_r$ are the associated parameters.

The proof of Theorem 2 is based on a basic property of the regression parameters $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b)$ of model (7), that is, that they are identical to log-linear parameters defined in selected marginal tables. Specifically, each $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b)$ coincides with the vector of log-linear parameters $\boldsymbol{\lambda}_{Ab}^{AB}$ of order $A \cup b$ in the marginal table $A \cup \mathrm{pre}(T)$. See Lemma 2 in the Appendix.

Theorem 2 shows also that the chain graph model $\mathbf{P}_{\mathrm{MR}}(G)$ is defined by a set of linear restrictions on a multivariate logistic parametrization and thus is a curved exponential family.

Example 5. From Theorem 2, the chain graph model of Figure 1(b) is defined by the equations

$$\eta^{(1)}(k,l) = \beta_{\phi}^{(1)} + \beta_{4}^{(1)}(k), \qquad \eta^{(2)}(k,l) = \beta_{\phi}^{(2)}, \qquad \eta^{(3)}(k,l) = \beta_{\phi}^{(3)} + \beta_{5}^{(3)}(l),$$
$$\eta^{(12)}(k,l) = \beta_{\phi}^{(12)} + \beta_{4}^{(12)}(k), \qquad \eta^{(13)}(k,l) = 0, \qquad \eta^{(23)}(k,l) = \beta_{\phi}^{(23)} + \beta_{5}^{(23)}(l),$$
$$\eta^{(123)}(k,l) = \beta_{\phi}^{(123)} + \beta_{4}^{(123)}(k) + \beta_{5}^{(123)}(l) + \beta_{45}^{(123)}(k,l)$$

and by the multivariate logistic model formula

$$X_1 : X_4, \qquad X_2 : 1, \qquad X_3 : X_5, \qquad X_{12} : X_4, \qquad X_{13} : 0, \qquad X_{23} : X_5, \qquad X_{123} : X_4 * X_5.$$

Notice that the marginal logit of X_2 does not depend on the variables X_4, X_5. This is denoted by $X_2 : 1$. On the other hand, the missing edge $(1, 3)$ with associated independence $1 \perp\!\!\!\perp 3|4, 5$ implies that the bivariate logit between X_1 and X_3 is zero, denoted by model formula $X_{13} : 0$. The above equations reflect exactly the independence structure encoded by the multivariate regression Markov property but leave a complete free model for the three-variable logistic parameter $\eta^{(123)}$.

Table 1 lists the parameters (and their log-linear interpretations) of the saturated model. The non-vanishing parameters of the chain graph model are in boldface. The shaded portion of the table indicates the interactions of an order higher than two. Therefore, the chain graph model contains seven parameters in the shaded area that have a more complex interpretation and that are not strictly needed to define the independence structure. This leads us to consider, as a starting model, a multivariate logistic regression model with no parameters of log-linear order higher than two and then use a backward selection strategy to test for the independencies. Some adjustment of the procedure is needed to include selected higher order interactions when needed. Notice also that the parameters in Table 1 form a marginal log-linear parametrization in the sense of Bergsma and Rudas [3], a result that can be proved for any discrete multivariate regression chain model. For an example see [19].

Table 1. Marginal log-linear parameters of the saturated model for a discrete multivariate logistic model with three responses and two explanatory variables. Each row lists log-linear parameters defined within a marginal table indicated in the last column. The non-zero terms of the chain graph model of Example 5 are shown in boldface. The shaded part of the table collects the interactions of an order higher than two

Logit	Parameters				Margin
	Const.	4	5	45	
1	**1**	**14**	15	145	145
2	**2**	24	25	245	245
3	**3**	34	**35**	345	345
12	**12**	**124**	125	1245	1245
13	13	134	135	1345	1345
23	**23**	234	**235**	2345	2345
123	**123**	**1234**	**1235**	**12345**	**12345**

A parallel multivariate logistic parametrization for the model $\mathbf{P}_{\mathrm{IV}}(G)$ can be obtained from Definition 2 and the associated characterization in terms of densities of Lemma 1 in the Appendix. In this case, using the factorization (3), the multivariate logistic models can be defined in the lower-dimensional conditional distributions $p_{T|\,\mathrm{pa}_D(T)}$. Therefore we state the following corollary.

Corollary 1. *The joint probability distribution of the random vector* $\mathbf{X}$ *belongs to* $\mathbf{P}_{\mathrm{IV}}(G)$ *if and only if it factorizes according to equation* (3), *and for each conditional distribution* $p(\mathbf{i}_T|\mathbf{i}_{\mathrm{pa}_D(T)})$, *for* $T \in \mathcal{T}$, *the multivariate logistic parameters are*

$$\boldsymbol{\eta}^{(A)}\big(\mathbf{i}_{\mathrm{pa}_D(T)}\big) = \begin{cases} \displaystyle\sum_{b\subseteq \mathrm{pa}_G(A)} \boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b) & \text{for all connected } A \subseteq T, \\ \mathbf{0} & \text{for all disconnected } A \subseteq T. \end{cases} \tag{13}$$

In the class of models defined in Remark 3, corresponding exactly to the factorization (3), all the independencies are obtained by setting $\mathrm{pa}_G(A) = \mathrm{pa}_D(T)$ for all $A \subseteq T$ in equation (11a).

4.2. Likelihood inference

The estimation of discrete multivariate regression chain models can be carried out by fitting separate multivariate logistic regression models to each factor $p_{T|\,\mathrm{pre}(T)}$ of the decomposition (4). Specifically, given a block T of responses and the group of covariates $\mathrm{pre}(T)$, we consider the table of frequencies $\mathbf{Y}_k$ for each covariate class k, where $k = 1, \ldots, K$ is an index numbering the cells of the marginal table $\mathcal{I}_{\mathrm{pre}(T)}$. Then we assume that each $\mathbf{Y}_k \sim M(n_k, \mathbf{p}_k)$ is multinomial with $\mathbf{p}_k = \mathbf{p}(\mathbf{i}_{\mathrm{pre}(T)})$. Given K independent observations $(\mathbf{Y}_1, n_1), \ldots, (\mathbf{Y}_K, n_K)$ the vector $\mathbf{Y} = \mathrm{vec}(\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ has a product-multinomial distribution and the log-likelihood is

$$l(\boldsymbol{\omega}) = \mathbf{y}^T\boldsymbol{\omega} - \mathbf{1}^T \exp(\boldsymbol{\omega}), \tag{14}$$

where $\boldsymbol{\omega} = \log E(\mathbf{Y}) = \log \boldsymbol{\mu}$ and $\mathbf{C}_T \log(\mathbf{M}_T\boldsymbol{\mu}) = \mathbf{Z}_r\boldsymbol{\beta}_r$, from (12). The maximization of this likelihood under the above linear constraints has been discussed by several authors; see [2,3,13, 15], among others.

Example 6. We give a simple illustration based on an application to data from the US General Social Survey [7], for years 1972–2006. The data are collected on 13 067 individuals on 5 variables. There are three binary responses concerning individual opinions (1 = favor, 2 = oppose) on legal abortion if pregnant as a result of rape, A; on death penalty for those convicted of murder, C; and on the introduction of police permits for buying guns, G. The potentially explanatory variables considered are J, job satisfaction (with three levels: 1 = very satisfied, 2 = moderately satisfied, 3 = a little or very dissatisfied), and S, gender (1 = male, 2 = female). We can interpret responses G and C as indicators of the attitude towards individual safety, while C and A are indicators of the concern for the value of human life, even in extreme situations.

The two explanatory variables turned out to be independent (with a likelihood ratio test statistic of $w = 0.79$, 1 d.f.). Hence, we concentrate on the model for the conditional distribution

Table 2. Multivariate regression chain graph model selection for GSS data. Model (1) is the pure independence model of Figure 3 for $p_{GCA|JS}$. Models (2)–(7) are fitted during the suggested model selection procedure. On the right are the fitted parameters for the best selected model

Model for $p_{GCA\|JS}$	Deviance	d.f.	Logit	Const.	J_{mdr}	J_{full}	S_{f}
(1) $G \perp\!\!\!\perp A\|J, S$ and $G \perp\!\!\!\perp J\|S$	12.84	10	G	0.766			0.766
(2) No 5-factor interaction	0.49	2	C	1.051	0.150	0.257	−0.458
(3) + no 4-factor interactions	5.59	11	A	1.826	−0.033	−0.245	−0.172
(4) + no 3-factor interactions	30.16	27	GC	−0.303			
(6) + Delete edge GA	33.38	28	CA	0.557			
(7) + Delete edge GJ	34.25	30					

$p_{GCA|JS}$. Here the saturated model (9) has 42 parameters and the structure of the parameters is that of Table 1, with the only modification of the dimensions of the interaction parameters involving the factor J, with three levels. We describe a hierarchical backward selection strategy. For this, we examine first the sequence of models obtained by successively removing the higher order interactions; see Table 1. Then we drop some of the remaining terms to fit independencies.

The results are shown in Table 2. The model with no interactions of an order higher than three has a deviance of 30.16 with 27 degrees of freedom adequate. From the edge exclusion deviances, we verify that we can remove the edges GA ($w = 33.38 - 30.16 = 3.22$, 1 d.f.) and GJ ($w = 34.25 - 33.38 = 0.87$, 2 d.f.). The final multivariate regression chain graph model, as shown in Figure 3(a), has a combined deviance of $34.25 + 0.79 = 35.04$ on 32 degrees of freedom.

Notice that the model includes independence and non-independence constraints, the latter following our preference for a model with all interpretable parameters. The chain graph model corresponding exactly to the implied independencies has far more parameters, with a deviance of $12.84 + 0.79 = 13.63$ against 12 degrees of freedom. While this model is adequate, the chosen model has a simpler interpretation. The fitted parameters are shown in Table 2 on the right. The first three rows give the parameters of three univariate logit regressions for being in favor of the issue. J_{mdr}, J_{full} measure the effect of moderate and full job satisfaction, respectively, with respect to a baseline level of no satisfaction, and S_{f} is the effect of females. Thus the effect

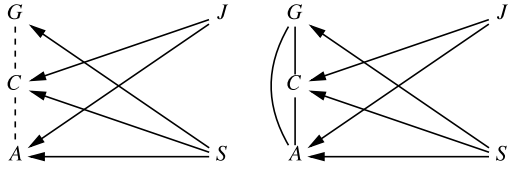

Figure 3. (a) The multivariate regression chain graph model fitted to GSS data (Deviance = 13.63, d.f. = 12). The final fitted model including further non-independence constraints has a Deviance = 35.04 on 32 d.f. (b) the best fitting LWF chain graph model (Deviance = 12.81, d.f. = 18).

of increased job satisfaction whatever the gender, is to increase the probability of being in favor of capital punishment and against abortion. Women are more favorable than males toward gun regulation and are more against the death penalty and abortion, all things being equal. The negative residual association between G and C and the positive one between C and A having accounted for gender and job satisfaction are as expected. As a comparison, in this example, a best-fitting classical chain graph model with LWF interpretation has one additional edge, as shown in Figure 3. The multivariate regression chain graph has a simpler interpretation in terms of three additive logistic regressions and two residual associations interpretable as deriving from two latent variables.

Appendix: Proofs

We shall assume for the joint distribution the existence of a density with respect to a product measure. Proofs using only basic properties of conditional independence can also be given, but are omitted for brevity.

Lemma 1. *The block-recursive Markov property of type IV is equivalent to the following three statements*: *for all* $T \in \mathcal{T}$

$$p_{T|\,\mathrm{pre}(T)} = p_{T|\,\mathrm{pa}_D(T)}, \tag{15a}$$

$$p_{A|\,\mathrm{pa}_D(T)} = p_{A|\,\mathrm{pa}_G(A)} \qquad \textit{for all connected } A \subseteq T, \tag{15b}$$

$$p_{A|\,\mathrm{pa}_D(T)} = \prod_j p_{A_j|\,\mathrm{pa}_D(T)} \qquad \textit{for all disconnected } A \subseteq T, \tag{15c}$$

where A_j, $j = 1, \ldots, r$, *are the connected components of* A, *if disconnected.*

Proof. Condition (IV0) states that the joint probability distribution obeys the *local directed Markov property* relative to the directed graph D of the chain components. Then, using the equivalence of the local and well-ordered local Markov property in directed graphs applied to the graph of the components as discussed in [9], Appendix A, (IV0) turns out to be equivalent to (15a) for any ordering of the components consistent with the chain graph. Moreover, condition (IV2) has been proved by [11] to be equivalent to the joint independence (15c). Statement (IV1) implies (15b) but it can be restricted to connected subsets A because, for disconnected subsets, it follows from (15c) and from (15b) restricted to connected sets. If A is disconnected, (15c) implies

$$p_{A|\,\mathrm{pa}_D(T)} = \prod_j p_{A_j|\,\mathrm{pa}_D(T)} = \prod_j p_{A_j|\,\mathrm{pa}_G(A_j)} = \prod_j p_{A_j|\,\mathrm{pa}_G(A)} \tag{16}$$

by applying (15b) to the connected sets A_j and noting that $\mathrm{pa}_G(A_j) \subseteq \mathrm{pa}_G(A)$. Therefore, $p_{A|\,\mathrm{pa}_D(T)} = p_{A|\,\mathrm{pa}_G(A)}$ and equation (IV1) follows. □

Then we are ready to prove that the multivariate regression Markov property is equivalent to the above block-recursive Markov property.

Proof of Theorem 1. We establish the equivalence of (1a) and (1b) with (15a), (15b) and (15c) of Lemma 1.

(Definition 1 implies Definition 2.) Equation (1a) implies $p_{A|\mathrm{pre}(T)} = p_{A|\mathrm{pa}_D(T)}$ for all connected A because $\mathrm{pa}_G(A) \subseteq \mathrm{pa}_D(T)$. Thus (1a) implies (15b) and (15a) for $A = T$, because any G_T is connected, by definition. Thus, if A is disconnected, (1b) gives

$$p_{A|\mathrm{pre}(T)} = \prod_j p_{A_j|\mathrm{pre}(T)} = \prod_j p_{A_j|\mathrm{pa}_D(T)} = p_{A|\mathrm{pa}_D(A)}$$

and (15c) follows.

(Definition 2 implies Definition 1.) Statement (15a) implies, for $A \subseteq T$, that $p_{A|\mathrm{pre}(T)} = p_{A|\mathrm{pa}_D(T)}$. Thus for all connected A, (15b) implies $p_{A|\mathrm{pre}(T)} = p_{A|\mathrm{pa}_G(A)}$, i.e., (1a). Moreover, if $A \subseteq T$ is disconnected, (15c) implies

$$p_{A|\mathrm{pre}(T)} = p_{A|\mathrm{pa}_D(T)} = \prod_j p_{A_j|\mathrm{pa}_D(T)} = \prod_j p_{A_j|\mathrm{pre}_D(T)},$$

that is, (1b). □

Given a subvector $\mathbf{X}_M$ of the given random vector $\mathbf{X}$, the log-linear expansion of its marginal probability distribution p_M is

$$\log p_M(\mathbf{i}_M) = \sum_{s \subseteq M} \lambda_s^M(\mathbf{i}_s), \tag{17}$$

where $\lambda_s^M(\mathbf{i}_s)$ defines the 'interaction' parameters of order $|s|$ in the baseline parametrization, that is, with the implicit constraint that the function returns zero whenever at least one of the indices in $\mathbf{i}_s$ takes the first level.

Lemma 2. *If $\eta^{(A)}(\mathbf{i}_A^\star|\mathbf{i}_{\mathrm{pre}(T)})$ is the multivariate logistic contrast of the conditional probability distribution $p_{A|\mathrm{pre}(T)}$ for A subset of T, then, with $B = \mathrm{pre}(T)$,*

$$\eta^{(A)}(\mathbf{i}_A^\star|\mathbf{i}_B) = \sum_{b \subseteq B} \lambda_{Ab}^{AB}(\mathbf{i}_A^\star, \mathbf{i}_b), \tag{18}$$

where $\lambda_{Ab}^{AB}(\mathbf{i}_A^\star, \mathbf{i}_b)$ are the log-linear interaction parameters of order $A \cup b$ in the marginal probability distribution p_{AB}.

Proof. First note that the multivariate logistic contrasts $\eta^{(A|B)}(\mathbf{i}_A^\star|\mathbf{i}_B)$ can be written

$$\eta^{(A|B)}(\mathbf{i}_A^\star|\mathbf{i}_B) = \sum_{s \subseteq A} (-1)^{|A \setminus s|} \log p_{AB}(\mathbf{i}_s^\star, \mathbf{i}_B, \mathbf{1}_{A \setminus s}). \tag{19}$$

Then we express the logarithm of the joint probabilities p_{AB} as the sum of log-linear interactions using (17),

$$\log p_{AB}(\mathbf{i}_s^\star, \mathbf{i}_B, \mathbf{1}_{A\setminus s}) = \sum_{a\subseteq A}\sum_{b\subseteq B}\lambda_{ab}^{AB}(\mathbf{i}_{a\cap s}^\star, \mathbf{1}_{a\setminus s}, \mathbf{i}_b) = \sum_{a\subseteq s}\sum_{b\subseteq B}\lambda_{ab}^{AB}(\mathbf{i}_a^\star, \mathbf{i}_b).$$

Therefore, by substitution into equation (19) we get

$$\begin{aligned}\eta^{A|B}(\mathbf{i}_A^\star|\mathbf{i}_B) &= \sum_{s\subseteq A}(-1)^{|A\setminus s|}\sum_{a\subseteq s}\sum_{b\subseteq B}\lambda_{ab}^{AB}(\mathbf{i}_a^\star, \mathbf{i}_b)\\ &= \sum_{b\subseteq B}\sum_{s\subseteq A}(-1)^{|A\setminus s|}\sum_{a\subseteq s}\lambda_{ab}^{AB}(\mathbf{i}_a^\star, \mathbf{i}_b) = \sum_{b\subseteq B}\lambda_{Ab}^{AB}(\mathbf{i}_A^\star, \mathbf{i}_b),\end{aligned}$$

where the last equality is obtained using a Möbius inversion; see [16], Lemma A.2, page 239. □

Lemma 2 is used in the proof of Theorem 2 given below.

Proof of Theorem 2. If (11a) holds for any chain component T, then for any connected set $A \subseteq T$, $\boldsymbol{\eta}^{(A)}(\mathbf{i}_{\mathrm{pre}(T)})$ is a function of $\mathbf{i}_{\mathrm{pa}_G(T)}$ only. Therefore, using the diffeomorphism and the property of upward compatibility discussed in Remark 5, the conditional distribution $p_{A|\,\mathrm{pre}(T)}$ coincides with $p_{A|\,\mathrm{pa}_G(A)}$ and condition (MR1) holds.

Conversely, if condition (MR1) holds and $p_{A|\,\mathrm{pre}(T)} = p_{A|\,\mathrm{pa}_G(A)}$, for all connected subsets A of T, then the components of $\boldsymbol{\eta}^{(A)}(\mathbf{i}_{\mathrm{pre}(T)})$ are

$$\begin{aligned}\eta^{(A)}\big(\mathbf{i}_A^\star|\mathbf{i}_{\mathrm{pre}(T)}\big) &= \sum_{s\subseteq A}(-1)^{|A\setminus s|}\log p\big(\mathbf{i}_s^\star, \mathbf{1}_{A\setminus s}\mid \mathbf{i}_{\mathrm{pa}_G(T)}\big)\\ &= \sum_{b\subseteq B}\lambda_{Ab}^{AB}(\mathbf{i}_A^\star, \mathbf{i}_b), \qquad \text{with } B = \mathrm{pa}_G(T)\end{aligned}$$

by Lemma 2, and thus (11a) holds with $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b) = \boldsymbol{\lambda}_{Ab}^{AB}(\mathbf{i}_b)$, where $\boldsymbol{\lambda}_{Ab}^{AB}(\mathbf{i}_b)$ denotes the vector of log-linear parameters $\lambda_{Ab}^{AB}(\mathbf{i}_A^\star, \mathbf{i}_b)$ for all $\mathbf{i}_A^\star \in \mathcal{I}_A^\star$.

Condition (MR2) of Definition 1 is equivalent to imposing that, for any chain component T, the conditional distribution $p_{T|\,\mathrm{pre}(T)}$ satisfies the independence model of a covariance subgraph G_T. In [14] and [18] it is proved that, given a joint distribution p_T, a covariance graph model is satisfied if and only if, in the multivariate logistic parameterization $\boldsymbol{\eta}_T$, $\boldsymbol{\eta}^{(A)} = \mathbf{0}$ for all disconnected sets $A \subseteq T$. Therefore, extending this result to the conditional distribution $p_{T|\,\mathrm{pre}(T)}$ and considering the diffeomorphism (7), condition (MR2) holds if and only if $\boldsymbol{\eta}^{(A)}(\mathbf{i}_B) = \mathbf{0}$ for every disconnected set $A \subseteq T$. Following the factorial model (7), $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b) = \mathbf{0}$ with $b \subseteq \mathrm{pre}(T)$ for each disconnected subset A of T. Notice that, by Lemma 2, $\boldsymbol{\beta}_b^{(A)}(\mathbf{i}_b) = \boldsymbol{\lambda}_{Ab}^{AB}(\mathbf{i}_b) = \mathbf{0}$, with $b \subseteq \mathrm{pre}(T)$. □

Acknowledgements

We thank Nanny Wermuth, D. R. Cox and two referees for helpful comments. This work was supported by MIUR, Rome, under the project PRIN 2007, XECZ7L001/2.

References

[1] Andersson, S., Madigan, D. and Perlman, M. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28** 33–85. MR1844349

[2] Bartolucci, F., Colombi, R. and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statist. Sinica* **17** 691–711. MR2398430

[3] Bergsma, W.P. and Rudas, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159. MR1892659

[4] Blomeyer, D., Laucht, M., Coneus, K. and Pfeiffer, F. (2009). Initial risk matrix, home resources, ability development, and children's achievment. *J. Eur. Econom. Assoc.* **7** 638–648.

[5] Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* **8** 204–218, 247–277. MR1243593

[6] Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies*. London: Chapman and Hall. MR1456990

[7] Davis, J., Smith, T. and Marsden, J.A. (2007). *General Social Surveys Cumulative Codebook: 1972–2006*. Chicago: NORC.

[8] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15** 736–753. MR2555197

[9] Drton, M. and Perlman, M.D. (2008). A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference* **138** 1179–1200. MR2416875

[10] Drton, M. and Richardson, T.S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regression models. *Biometrika* **91** 383–392. MR2081308

[11] Drton, M. and Richardson, T.S. (2008). Binary models for marginal independence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 287–309. MR2424754

[12] Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.* **17** 333–353. MR1096723

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

[14] Kauermann, G. (1995). A note on multivariate logistic models for contingency tables. *Austral. J. Statist.* **39** 261–276. MR1616070

[15] Lang, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.* **24** 726–752. MR1394985

[16] Lauritzen, S.L. (1996). *Graphical Models*. Oxford: Oxford Univ. Press. MR1419991

[17] Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57. MR0981437

[18] Lupparelli, M., Marchetti, G.M. and Bergsma, W.P. (2009). Parameterizations and fitting of bi-directed graph models to categorical data. *Scand. J. Statist.* **36** 559–576. MR2549710

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

[20] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall. MR0727836

[21] Richardson, T.S. (2003). Markov property for acyclic directed mixed graphs. *Scand. J. Statist.* **30** 145–157. MR1963898

[22] Studený, M. (2005). *Probabilistic Conditional Independence Structures*. London: Springer.

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

[24] Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 687–717. MR2088296

[25] Wilkinson, G.N. and Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *J. Roy. Statist. Soc. Ser. C* **22** 392–399.

Received June 2009 and revised April 2010