Statistica Sinica **17**(2007), 691-711

AN EXTENDED CLASS OF MARGINAL LINK FUNCTIONS FOR MODELLING CONTINGENCY TABLES BY EQUALITY AND INEQUALITY CONSTRAINTS

Francesco Bartolucci[1], Roberto Colombi[2] and Antonio Forcina[1]

[1] *Università di Perugia and* [2] *Università di Bergamo*

Abstract: We extend Bergsma and Rudas (2002)'s hierarchical complete marginal parameterization to allow for logits and higher order effects of global and continuation type which may be more suitable with ordinal data. We introduce a general definition of marginal interaction parameters and show that this parameterization constitutes a link function so that linear models defined by equality and inequality constraints may be fitted and tested by extending the methods of Colombi and Forcina (2001). Computation and asymptotic properties of maximum likelihood estimators are discussed, and the asymptotic distribution of the likelihood ratio test is derived.

Key words and phrases: Chi-bar-squared distribution, likelihood inference, marginal models, one-sided linear hypothesis.

1. Introduction

Marginal models for the analysis of frequency tables have been developed during the last decade in response to the need to overcome certain limitations of log-linear models. One limitation is that lower order log-linear effects do not describe the marginal distribution to which they refer. For instance, main effects defined as the logits of the variable of interest averaged across all the possible configurations of the remaining variables may differ substantially from the corresponding marginal logits with which they are sometimes confused. The other limitation is that, for those variables which have an ordinal nature, different types of logit (and similar higher order effects) based on the cumulative distribution function or the survival function may be more meaningful; these logits, known as global, continuation or reverse continuation, are not log-linear parameters.

Molenberghs and Lesaffre (1994) studied a class of regression models where the univariate marginal logits and log-odds ratios of global type are allowed to depend on covariates. This approach was generalized by the multivariate logistic transform of Glonek and McCullagh (1995), who proposed a composite link function whose elements are the highest order interactions that can be defined

within each possible marginal distribution of the response variables; they also allowed parameters based on global logits for ordinal variables. Lang and Agresti (1994) and Lang (1996) studied a class of marginal models which is completely general but, because of this, also less structured. To bypass the fact that models of conditional independence cannot be expressed as linear constraints on the parameters of the multivariate logistic transform, Glonek (1996) introduced a *hybrid* parameterization which combines the multivariate logistic transform within low order marginals with log-linear parameters for the remaining higher order effects; parametrizations of this type were also used by Fitzmaurice and Laird (1993) and Fitzmaurice, Laird and Rotnitzky (1993) within regression models for the analysis of longitudinal data. This approach was further extended by Bergsma and Rudas (2002) by allowing additional log-linear parameters within a selected set of marginal distributions of interest.

Models involving linear inequality constraints arise naturally with ordinal variables and are needed to define hypotheses of stochastic dominance (Dardanoni and Forcina (1998)), or various notions of positive dependence (Bartolucci, Forcina and Dardanoni (2001)). A very general class of models defined by equality and inequality constraints on marginal parameters similar to those introduced by Glonek, has been studied by Colombi and Forcina (2001).

In the present paper we extend the marginal parameterization introduced by Bergsma and Rudas (2002) to allow for logits (and higher order effects) of different types and extend the results of Colombi and Forcina (2001) for computing maximum likelihood estimates under linear equality and inequality constraints, and for constructing an analysis of deviance table. In Section 2 we propose a definition of marginal interaction parameters of a general type (local, global, continuation) and study some of their properties. This throws new light on the interpretation of parameters and on the connection between hierarchical complete marginal parameterizations and block-recursive models. The main results are contained in Section 3, where we prove that the proposed parameterization defines a link function and we indicate when its elements are variation independent; our proof combines Bergsma and Rudas' idea of the recursive use of a mixed parameterization (Barndorff-Nielsen (1978)) with a new tool based on partially dichotomized tables. Computation and asymptotic properties of the maximum likelihood estimator are discussed in Section 4.

2. Generalized Marginal Interactions

This paper is about a new class of models for the joint distribution of a set of categorical response variables, conditional on a set of discrete explanatory variables. However as the main issues may be discussed conditionally on a given configuration of the explanatory variables, explicit reference to it will be omitted.

More formally, consider the joint distribution of q response variables $B_1, \ldots, B_q$, with B_j taking values in $\{1, \ldots, b_j\}$; this distribution identifies a contingency table having $t = \prod_1^q b_j$ cells. To be concise, the set of response variables that defines a given marginal distribution will be denoted by the set of indices of the corresponding variables and $\mathcal{Q} = \{1, \ldots, q\}$ will refer to the full joint distribution. The vector containing the cell probabilities of such a distribution, ordered lexicographically, will be denoted by $\boldsymbol{\pi}$.

Let Π denote the t-dimensional simplex: $\{\boldsymbol{\pi} : \pi_i > 0, \sum_1^t \pi_i = 1\}$. Formally, any invertible mapping $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\pi})$: $\Pi \to \Omega \subseteq \Re^{t-1}$ defines a *parameterization* and the elements of the vector $\boldsymbol{\eta}$ will be called *parameters*. In practice, we are interested in parameters which describe relevant aspects of the joint distribution and are such that hypotheses of interest may be expressed by linear constraints. Parameters defined as contrasts among the logarithms of probabilities of disjoint subsets of cells will be called *interactions*. As in Bergsma and Rudas (2002) (hereafter BR for brevity), an interaction will be characterized by the set $\mathcal{I}$ of variables involved and the marginal distribution $\mathcal{M}$ where it is defined. A general definition of interaction parameters will be formulated in 2.1 and some properties of linear transformations of these interactions will be examined in 2.2; this will be used to motivate and extend BR's definition of marginal parameterizations in 2.3. The aim of the final subsection, 2.4, is to provide additional motivation for BR's approach to allocating interactions within marginal distributions.

2.1. Generalized interaction parameters

Logits are the most elementary interaction parameters defined by contrasts between two disjoint subsets of cells within a univariate marginal (or conditional) distribution. Logits of four different types have been used in the literature: *local* (l), *global* (g), *continuation* (c) and *reverse continuation* (r). Colombi and Forcina (2001) discuss the interpretation of these logits; essentially, logits of type g, c and r may be used only when categories follow a natural order. Since logits of type r may be obtained from logits of type c when the order of categories is reversed, logits of this type need not be examined explicitly. It is well known that only logits of type l are log-linear within the corresponding marginal. Clearly, the type of logits should be chosen so as to suit the nature of each response variable. Note, however, that while it would be inappropriate to use logits of type g or c when categories do not follow a natural order, it makes sense to use logits of type l with both ordinal and non-ordinal variables. Different *types of contrast* can be used to define ordinary log-linear interactions; here we adopt contrasts between adjacent categories because this approach extends naturally to logits of global or continuation type.

We define below a general class of interaction parameters which includes the four types of logit mentioned above and the 16 types of log-odds ratio discussed by Douglas et al. (1990) as measures of bivariate association. Our approach, which extends naturally to interactions of any order, is based on the idea that the kind of dichotomy implied by the type of logit adopted for each variable should carry over when defining higher order interactions within the same marginal distribution. As we explain below it may be convenient to allow that interactions defined within different marginals are based on different types of logit for the same response variable.

Within a given marginal $\mathcal{M}$, assume that each response variable B_j, $j \in \mathcal{M}$, is assigned a given type of logit. For any *cut point* $x_j < b_j$, define the event $\mathcal{B}(x_j, 0)$ to be equal to $\{x_j\}$ if the logit is of local or continuation type and to $\{1, \ldots, x_j\}$ for global logits; similarly, the event $\mathcal{B}(x_j, 1)$ is equal to $\{x_j + 1\}$ if the logit is of local type and to $\{x_j + 1, \ldots, b_j\}$ for global or continuation logits. Lastly, define the marginal probabilities

$$p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{M}}) = p(B_j \in \mathcal{B}(x_j, h_j),\ \forall j \in \mathcal{M}),$$

where $\boldsymbol{x}_{\mathcal{M}}$ is a row vector of cut points x_j, $j \in \mathcal{M}$, and $\boldsymbol{h}_{\mathcal{M}}$ is a row vector whose elements, h_j, $j \in \mathcal{M}$, are equal to zero or to one. A generalized interaction for the variables in $\mathcal{I}$, computed within the marginal $\mathcal{M}$, is defined through

$$\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{I}}) = \sum_{\mathcal{J}\subseteq\mathcal{I}} (-1)^{|\mathcal{I}\setminus\mathcal{J}|} \log p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{I}}, \boldsymbol{0}_{\mathcal{I}\setminus\mathcal{J}}, \boldsymbol{1}_{\mathcal{J}}), \quad (1)$$

where the binary vector $\boldsymbol{h}_{\mathcal{M}}$ has been split into three components, and $\boldsymbol{1}_{\mathcal{J}}$ denotes a vector of $|\mathcal{J}|$ ones. These parameters may be interpreted as log-linear contrasts computed within the marginal table obtained by dichotomizing the original variables according to their respective types of logit. Note that (1) is defined also for $\mathcal{I} = \emptyset$, in which case the corresponding interaction parameter equals $\log p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{M}})$. The recursive nature of this definition is made explicit in the following.

Proposition 1. *For any $\mathcal{H} \subseteq \mathcal{M}\setminus\mathcal{I}$,* (1) *is equivalent to*

$$\eta_{\mathcal{I}\cup\mathcal{H};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{H}} \mid \boldsymbol{x}_{\mathcal{L}}; \boldsymbol{h}_{\mathcal{L}}) = \sum_{\mathcal{K}\subseteq\mathcal{H}} (-1)^{|\mathcal{H}\setminus\mathcal{K}|} \eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \boldsymbol{h}_{\mathcal{L}}, \boldsymbol{0}_{\mathcal{H}\setminus\mathcal{K}}, \boldsymbol{1}_{\mathcal{K}}), \quad (2)$$

where $\mathcal{L} = \mathcal{M}\setminus(\mathcal{I}\cup\mathcal{H})$.

Proof of Proposition 1. Expand the left-hand side of (2) as in (1) and note that, since $\mathcal{I}$ and $\mathcal{H}$ are disjoint, any subset of $\mathcal{I}\cup\mathcal{H}$ may be expressed as $\mathcal{K}\cup\mathcal{J}$,

with $\mathcal{K} \subseteq \mathcal{H}$ and $\mathcal{J} \subseteq \mathcal{I}$, so that

$$\eta_{\mathcal{I}\cup\mathcal{H};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{H}} \mid \boldsymbol{x}_{\mathcal{L}}; \boldsymbol{h}_{\mathcal{L}}) = \sum_{\mathcal{K}\subseteq\mathcal{H}} (-1)^{|\mathcal{H}\setminus\mathcal{K}|} \sum_{\mathcal{J}\subseteq\mathcal{I}} (-1)^{|\mathcal{I}\setminus\mathcal{J}|} \log p_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}; \boldsymbol{h}_{\mathcal{L}}, \mathbf{0}_{(\mathcal{H}\cup\mathcal{I})\setminus(\mathcal{K}\cup\mathcal{J})}, \mathbf{1}_{\mathcal{K}\cup\mathcal{J}}).$$

For a given $\mathcal{K} \subseteq \mathcal{H}$, the binary vector $\boldsymbol{h}_{\mathcal{M}\setminus\mathcal{I}} = (\boldsymbol{h}_{\mathcal{L}}, \mathbf{0}_{\mathcal{H}\setminus\mathcal{K}}, \mathbf{1}_{\mathcal{K}})$ is fixed while the component $(\mathbf{0}_{\mathcal{I}\setminus\mathcal{J}}, \mathbf{1}_{\mathcal{J}})$ varies within the inner sum and this sum is equal to the right-hand side of (1), so that (2) follows by substitution.

In the special case where $\mathcal{H} = \{h\}$, if we write $\mathcal{G} = \mathcal{I} \cup \{h\}$, (2) becomes

$$\eta_{\mathcal{G};\mathcal{M}}(\boldsymbol{x}_{\mathcal{G}}|\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{G}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{G}}) = \eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \boldsymbol{h}_{\mathcal{L}}, 1) - \eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \boldsymbol{h}_{\mathcal{L}}, 0); \quad (3)$$

this recursive relation implies that any interaction parameter for variables in $\mathcal{G}$ may be interpreted as the difference between the corresponding interaction parameter involving $\mathcal{I} = \mathcal{G}\setminus\{h\}$ by changing the conditioning variable B_h from $\mathcal{B}(x_h, 0)$ to $\mathcal{B}(x_h, 1)$, no matter how we choose $\{h\} \in \mathcal{G}$.

2.2. Linear transformations of generalized interaction parameters

Interaction parameters, as defined in (1), depend on conditioning variables through their cut points $\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}$ and the binary vector $\boldsymbol{h}_{\mathcal{M}\setminus\mathcal{I}}$. Thus, when $\mathcal{M}\setminus\mathcal{I}$ is not empty, the same intuitive notion of interaction could be parameterized in many different ways. In this section, we indicate how such parameters are linearly related; these results indicate that we may limit our attention to the interaction parameters with $\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}} = \mathbf{1}_{\mathcal{M}\setminus\mathcal{I}}$ and $\boldsymbol{h}_{\mathcal{M}\setminus\mathcal{I}} = \mathbf{0}_{\mathcal{M}\setminus\mathcal{I}}$, that is to say to the interactions where the conditioning variables are fixed at their first category.

Now let $\boldsymbol{h}_{\mathcal{L}} = \mathbf{0}_{\mathcal{L}}$ in (2) and apply the Möbius Inversion Lemma (Lauritzen (1996, p.239)) to write $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \mathbf{0}_{\mathcal{L}}, \mathbf{1}_{\mathcal{H}})$ as a sum of higher order interactions $\eta_{\mathcal{I}\cup\mathcal{K};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{K}} \mid \boldsymbol{x}_{\mathcal{M}\setminus(\mathcal{I}\cup\mathcal{K})}; \mathbf{0}_{\mathcal{M}\setminus(\mathcal{I}\cup\mathcal{K})})$:

$$\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \mathbf{0}_{\mathcal{L}}, \mathbf{1}_{\mathcal{H}}) = \sum_{\mathcal{K}\subseteq\mathcal{H}} \eta_{\mathcal{I}\cup\mathcal{K};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{K}} \mid \boldsymbol{x}_{\mathcal{M}\setminus(\mathcal{I}\cup\mathcal{K})}; \mathbf{0}_{\mathcal{M}\setminus(\mathcal{I}\cup\mathcal{K})}). \quad (4)$$

This equation says that, once the interactions of type $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{J}}; \boldsymbol{h}_{\mathcal{M}\setminus\mathcal{J}})$ with $\boldsymbol{h}_{\mathcal{M}\setminus\mathcal{J}} = \mathbf{0}$ are known, any other generalized interaction may be reconstructed by linear transformation. In addition note that, when B_j has logits of local type, $\mathcal{B}(x_j, 0) = \mathcal{B}(x_j - 1, 1)$, so that if all variables in $\mathcal{M}\setminus\mathcal{I}$ have logits of local type, repeated use of (4) leads to an expression for $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\setminus\mathcal{I}}; \mathbf{0}_{\mathcal{L}}, \mathbf{1}_{\mathcal{H}})$ in terms of interactions of the form $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}} \mid \mathbf{1}_{\mathcal{M}\setminus\mathcal{J}}; \mathbf{0}_{\mathcal{M}\setminus\mathcal{J}})$, with $\mathcal{I} \subseteq \mathcal{J}$; from now on these interaction parameters will be simply denoted by $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}})$.

Before deriving the general expression for reducing any interaction parameter to a linear function of simpler interactions of the form $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}})$, we give an example which may help to illustrate the basic idea.

Example 2.1. Let $\mathcal{M} = \{1, 2, 3\}$ with B_2 and B_3 having logits of local type with four and two levels, respectively. By (4) we have $\eta_{\{1\};\mathcal{M}}(1 \mid (3\ 2); (0\ 1)) = \eta_{\{1\};\mathcal{M}}(1 \mid (3\ 2); (0\ 0)) + \eta_{\{1,3\};\mathcal{M}}((1\ 2) \mid 3; 0)$; the first term on the right-hand side is equal to $\eta_{\{1\};\mathcal{M}}(1 \mid (2\ 1); (1\ 1))$ which, in turn, is equal to $\eta_{\{1\};\mathcal{M}}(1) + \eta_{\{1,2\};\mathcal{M}}((1\,1)) + \eta_{\{1,2\};\mathcal{M}}((1\,2)) + \eta_{\{1,3\};\mathcal{M}}((1\,1)) + \eta_{\mathcal{M};\mathcal{M}}((1\,1\,1)) + \eta_{\mathcal{M};\mathcal{M}}((1\,2\,1))$, while the second term may be expanded as $\eta_{\{1,3\};\mathcal{M}}((1\ 2)) + \eta_{\mathcal{M};\mathcal{M}}((1\ 1\ 2)) + \eta_{\mathcal{M};\mathcal{M}}((1\ 2\ 2))$.

If we let $\mathcal{J}(1) = \{j \in \mathcal{M}\backslash(\mathcal{I} \cup \mathcal{K}) : x_j > 1\}$ and $\mathcal{G} = \mathcal{M}\backslash(\mathcal{I} \cup \mathcal{K} \cup \mathcal{J}(1))$, we can rewrite each term on the right-hand side of (4) as $\eta_{\mathcal{I}\cup\mathcal{K};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{K}} \mid \boldsymbol{x}_{\mathcal{G}}, \boldsymbol{x}_{\mathcal{J}(1)} - 1; \mathbf{0}_{\mathcal{G}}, \mathbf{1}_{\mathcal{J}(1)})$, which can be expanded again as in (4) and the process can be iterated until all terms on the right-hand side are in the form $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}})$, with $\mathcal{I} \subseteq \mathcal{J}$. To describe this process formally, let m be the value of the largest element of $\boldsymbol{x}_{\mathcal{M}\backslash\mathcal{I}}$ minus 1, and define $\mathcal{J}(h) = \{j \in \mathcal{M}\backslash(\mathcal{I} \cup \mathcal{K} \cup \mathcal{L}(1) \cup \cdots \cup \mathcal{L}(h-1)) : x_j > h\}$, $h = 2, \ldots m$, where $\mathcal{L}(h) \subseteq \mathcal{J}(h)$, $\bar{\mathcal{L}} = \mathcal{J}(1)\backslash\mathcal{L}(1)$ and $\mathcal{H}(h) = \mathcal{I} \cup \mathcal{K} \cup \mathcal{L}(1) \cup \cdots \cup \mathcal{L}(h)$. Then

$$\begin{aligned}
&\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\backslash\mathcal{I}}; \mathbf{0}_{\mathcal{L}}, \mathbf{1}_{\mathcal{H}}) \\
&\quad = \sum_{\mathcal{K}\subseteq\mathcal{H}} \sum_{\mathcal{L}(1)\subseteq\mathcal{J}(1)} \eta_{\mathcal{H}(1);\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{K}}, \boldsymbol{x}_{\mathcal{L}(1)} - 1 \mid \boldsymbol{x}_{\mathcal{G}}, \boldsymbol{x}_{\bar{\mathcal{L}}} - 1; \mathbf{0}_{\mathcal{M}\backslash\mathcal{H}(1)}) \\
&\quad = \sum_{\mathcal{K}\subseteq\mathcal{H}} \sum_{\mathcal{L}(1)\subseteq\mathcal{J}(1)} \cdots \sum_{\mathcal{L}(m)\subseteq\mathcal{J}(m)} \eta_{\mathcal{H}(m);\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}\cup\mathcal{K}}, \boldsymbol{x}_{\mathcal{L}(1)} - 1, \ldots, \boldsymbol{x}_{\mathcal{L}(m)} - m). \qquad (5)
\end{aligned}$$

Below we limit our attention only to interactions of the simple form $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}})$, where the conditioning variables are fixed at their initial level.

When the conditioning variables have logits of local type, (5) implies that any linear constraint on the generalized interactions $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}} \mid \boldsymbol{x}_{\mathcal{M}\backslash\mathcal{I}}; \boldsymbol{h}_{\mathcal{M}\backslash\mathcal{I}})$ may be written as a linear constraint on $\eta_{\mathcal{J};\mathcal{M}}(\boldsymbol{x}_{\mathcal{J}})$, so there is no restriction in limiting our attention to these parameters. Setting the type of logits of variables in $\mathcal{M}\backslash\mathcal{I}$ to local within $\mathcal{M}$ makes sense in certain recursive models described toward the end of this section. When not all the variables in $\mathcal{M}\backslash\mathcal{I}$ have logits of local type, parameters defined by fixing the conditioning variables to a different reference category, or by averaging across all possible configurations of the conditioning variables, are no longer linearly related.

2.3. Marginal parameterizations

We now examine the issue of allocating interaction parameters among the marginals within which they may be defined. In doing so we extend BR's notion

of *complete hierarchical parameterization* to the generalized interaction parameters $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}})$; we also rephrase their formulation in a way which is useful for the following treatment and provide some additional motivation. Denote by $\mathcal{M}_1, \ldots, \mathcal{M}_s$ an ordered set of marginals of interest, and by $\mathcal{F}_m$ the collection of the sets $\mathcal{I}$ such that $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}})$ is defined within $\mathcal{M}_m$. Also let $\mathcal{P}(\mathcal{J})$ be the set of all non-empty subsets of $\mathcal{J}$ and $\mathcal{P}_m$ be a short-hand notation for $\mathcal{P}(\mathcal{M}_m)$.

Definition 1. A marginal parameterization is called complete and hierarchical if (i) the sequence of marginals $\mathcal{M}_1, \ldots, \mathcal{M}_s$ is non-decreasing and $\mathcal{M}_s = \mathcal{Q}$, (ii) $\mathcal{F}_1 = \mathcal{P}_1$ and $\mathcal{F}_m = \mathcal{P}_m \setminus \bigcup_1^{m-1} \mathcal{F}_h$ for $m > 1$.

The definition implies that any interaction with $\mathcal{I} \in \mathcal{P}(\mathcal{Q})$ is defined in one and only one marginal distribution $\mathcal{M}_m$, a feature which is called *completeness*. The definition also implies that $\mathcal{F}_m$ cannot be empty and constitutes an *ascending* class of subsets of $\mathcal{P}_m$, while its complement with respect to $\mathcal{P}_m$, which we denote by $\mathcal{R}_m$, is a *descending* class in BR's terminology. Ascending means that, if a subset of $\mathcal{P}_m$ belongs to $\mathcal{F}_m$, the same is true for any larger subset. Because of this latter property, the parameterization is called *hierarchical*; in practice each interaction is defined within the first marginal within which it is contained. We limit our attention below to parameterizations which are complete and hierarchical. If a parameterization was complete but one or more $\mathcal{F}_m$ were not ascending classes, (5) could no longer be applied. This means that if, for instance, the interactions for $\mathcal{I} = \{1, 2, 3\}$ are defined within a marginal larger than $\mathcal{M} = \{1, 2, 3\}$, interactions defined within $\mathcal{M}$ by fixing the conditioning variables to a reference category, or by averaging with respect to all possible configurations, are no longer linearly related.

A complete and hierarchical set of generalized marginal interaction parameters has elements $\eta_{\mathcal{I};\mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}})$, for any $m = 1, \ldots, s$, any $\mathcal{I} \in \mathcal{F}_m$, and any configuration of cut points $\boldsymbol{x}_{\mathcal{I}} \in \prod_{j \in \mathcal{I}} \{1, \ldots, b_j - 1\}$; this set has $t - 1$ elements, as can be verified by ordinary calculation. These elements may be arranged into the vector $\boldsymbol{\eta}$ explicitly written in matrix form as

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

where the rows of the matrix $\boldsymbol{C}$ are contrasts, $\boldsymbol{M}$ is a matrix of zeros and ones which sums the probabilities of appropriate cells, and the $\log(\cdot)$ operator is coordinate-wise. A simple algorithm for constructing these matrices is given in the appendix (see also Colombi and Forcina (2001)). A given parameterization determines the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ and thus Ω, the range of values for $\boldsymbol{\eta}$ which correspond to a probability distribution $\boldsymbol{\pi}$. Within a given parameterization, a value $\boldsymbol{\eta} \in \Omega$ will be called *compatible*.

Following BR, we should also recall that the sequence $\mathcal{M}_1, \ldots, \mathcal{M}_s$ is *ordered decomposable* if the class of *maximal sets* within any subsequence $\mathcal{M}_1, \ldots, \mathcal{M}_m$, $m \geq 3$, is decomposable (Haberman (1974, p.166)), where a set is *maximal* if it is not contained in any other set. This notion is crucial for understanding when a marginal parameterization is *variation independent*, i.e., when $\Omega = \mathcal{R}^{t-1}$.

2.4. Block-recursive models

An important context where the flexibility allowed by complete hierarchical parameterizations can be exploited is when $\mathcal{Q}$ is partitioned into $\mathcal{U}_1, \ldots, \mathcal{U}_s$, so that, for any $m = 2, \ldots, s$, the variables in $\bigcup_1^{m-1} \mathcal{U}_h$ are potentially explanatory for the variables in $\mathcal{U}_m$. In this case a model with a *block-recursive* structure (Lauritzen (1996, Ch.4)) may be formulated by letting $\mathcal{M}_1 = \mathcal{U}_1$ and $\mathcal{M}_m = \mathcal{M}_{m-1} \cup \mathcal{U}_m$, for $m = 2, \ldots, s$, where the variables in $\mathcal{M}_{m-1}$ are assigned logits of local type within $\mathcal{M}_m$. It follows that, apart from $\mathcal{F}_1$ which equals $\mathcal{P}_1$, $\mathcal{F}_m$ contains all sets of the form $\mathcal{I} = \mathcal{K} \cup \mathcal{L}$, with $\mathcal{K} \in \mathcal{P}(\mathcal{U}_m)$ and $\mathcal{L} \in \mathcal{P}(\mathcal{M}_{m-1}) \cup \{\emptyset\}$. This formulation is such that all interaction parameters that define the conditional distributions of $\mathcal{U}_m$ given $\mathcal{M}_{m-1}$ are linear functions of interaction parameters as indicated by (5). For instance, the hypothesis that the variables in $\mathcal{U}_m$ are independent from, say, those in $\mathcal{M}_i$ given the variables in $\mathcal{M}_{m-1} \backslash \mathcal{M}_i$, may be formulated by constraining to 0 all $\eta_{\mathcal{I};\mathcal{M}_m}(\boldsymbol{x}_\mathcal{I})$ such that $\mathcal{I} \cap \mathcal{M}_i$ is non-empty. More sophisticated constraints on the conditional distribution of $\mathcal{U}_m$ given $\mathcal{M}_{m-1}$ (as, for instance, those implying that certain elements in $\mathcal{U}_m$ are positively associated with certain variables in previous blocks), may be expressed directly as the constraint that the corresponding conditional interactions given by (5) are non-negative. It is easily verified that the above sequence of marginals defines a parameterization which is complete, hierarchical and ordered decomposable.

A different approach for allocating interactions in a block-recursive context is as follows. Start with the same sequence of marginals as above and, after $\mathcal{M}_{m-1}$, insert the subsequence of marginals $\mathcal{M}_{ml} = \mathcal{K}_{ml} \cup \mathcal{M}_{m-1}$ where the sequence of subsets $\mathcal{K}_{ml}$, for all $\mathcal{K}_{ml} \in \mathcal{P}(\mathcal{U}_m)$, is non-decreasing so that the resulting parameterization is hierarchical. This approach, which combines the recursive structure with Glonek and McCullagh's multivariate logistic transform, may be desirable if we are interested in modelling constraints on various marginal distributions within $\mathcal{U}_m$ conditionally on $\mathcal{M}_{m-1}$. Notice that, for any $\mathcal{L} \in \mathcal{P}(\mathcal{M}_{m-1}) \cup \{\emptyset\}$, $\mathcal{L} \cup \mathcal{K}_{ml}$ is in $\mathcal{F}_{ml}$. Hence, if all variables in $\mathcal{M}_{m-1}$ have logits of local type within $\mathcal{M}_{ml}$, (5) can be used again to transform any constraint on the interactions in $\mathcal{K}_{ml}$, given any possible configuration of the variables in $\mathcal{M}_{m-1}$, into a constraint on our interaction parameters. For instance, to state that B_j, $j \in \mathcal{U}_m$, is independent of the variables in $\mathcal{M}_i$ given those in $\mathcal{M}_{m-1} \backslash \mathcal{M}_i$, all interaction parameters for sets of the form $\{j\} \cup \mathcal{L}$ such that $\mathcal{L} \cap \mathcal{M}_i$ is not empty must be

0. The assumption that B_i and B_j, $i, j \in \mathcal{U}_m$, are marginally independent, given variables in $\mathcal{M}_{m-1}$, requires that all interaction parameters for sets of the form $\{i\} \cup \{j\} \cup \mathcal{L}$, where $\mathcal{L}$ is either a subset of $\mathcal{M}_{m-1}$ or the empty set, be 0. Note that in both examples, all the interactions to be constrained are defined within the same marginal.

3. Properties of Marginal Link Functions

We now examine the class of parameterizations defined by (6) and show that it is invertible and has second order derivatives, so that, because of the similarity with generalized linear models, it may be called a *marginal link function*. We also provide conditions for the elements of the link function to be variation independent. Both issues are meant to extend BR's results to generalized interaction parameters. The most difficult step consists in showing that (6) defines a diffeomorphic transformation between Π and Ω.

As in BR, we exploit the fact that the probability density defined by the vector $\boldsymbol{\pi}$ is multinomial and thus belongs to the exponential family. Note that (see for example Bartolucci and Forcina (2002)) the relation between the vector of joint probabilities $\boldsymbol{\pi}$ and the corresponding vector of canonical parameters, denoted by $\boldsymbol{\lambda}$, is determined by a $t \times (t-1)$ matrix $\boldsymbol{G}$ which, apart from being of full rank and such that its column space does not contain the vector $\mathbf{1}$, may be arbitrary:

$$\log(\boldsymbol{\pi}) = \boldsymbol{G}\boldsymbol{\lambda} - \mathbf{1}\log[\mathbf{1}'\exp(\boldsymbol{G}\boldsymbol{\lambda})]. \tag{7}$$

Basically, this is a log-linear model with a scaling factor so that probabilities sum to 1. The inverse transformation may be written as $\boldsymbol{\lambda} = \boldsymbol{K}\log(\boldsymbol{\pi})$, where the matrix $\boldsymbol{K}$, given in the appendix, denotes the left inverse of $\boldsymbol{G}$ that is orthogonal to $\mathbf{1}$. Also let $\boldsymbol{\mu} = \boldsymbol{G}'\boldsymbol{\pi}$ denote the vector of *mean value parameters* and recall that, since the multinomial is a *regular* and *steep* exponential family, the mapping from $\boldsymbol{\lambda}$ to $\boldsymbol{\mu}$ is a diffeomorphism (Barndorff-Nielsen (1978, p.121)).

Since the design matrix may be arbitrary, there is no loss of generality in assuming that it has the following specific form, which allows substantial simplification. Let $\boldsymbol{G}_{\mathcal{I}}$ denote the block of columns of $\boldsymbol{G}$ that correspond to the log-linear interaction $\mathcal{I}$, and let

$$\boldsymbol{G}_{\mathcal{I}} = \bigotimes_{j=1}^{q} \boldsymbol{G}_{\mathcal{I},j}, \qquad \boldsymbol{G}_{\mathcal{I},j} = \begin{cases} \boldsymbol{T}_j & \text{if } j \in \mathcal{I} \\ \mathbf{1}_{b_j} & \text{otherwise} \end{cases},$$

where $\boldsymbol{T}_j$ is obtained from a $b_j \times b_j$ lower triangular matrix of ones by removing the first column. We show in the appendix that the canonical parameters corresponding to this design matrix coincide with the log-linear interactions of local type within the full joint distribution, and that the elements of $\boldsymbol{\mu}$ are the probabilities $\mu_{\mathcal{I}}(\boldsymbol{x}_{\mathcal{I}}) = p(B_j > x_j,\ \forall j \in \mathcal{I})$.

3.1. Partially dichotomized tables

We now introduce an auxiliary tool by means of which interaction parameters involving variables treated with logits of non-local type may be seen as interaction parameters of local type when these variables are dichotomized according to the type of logit. This tool, combined with the recursive notion of mixed parameterization of the exponential family, will be used to prove that (6) defines a *marginal link function*.

Within a given marginal $\mathcal{M}_m$, let $\mathcal{D}_m = \{j \in \mathcal{M}_m : B_j \text{ not of local type}\}$. For a vector of cut points $\boldsymbol{d}_{\mathcal{D}_m}$ with elements d_j, $\forall j \in \mathcal{D}_m$, define $pdt(m, \boldsymbol{d}_{\mathcal{D}_m})$ to be the collapsed table of dimension $\mid \mathcal{M}_m \mid$ and categories $\mathcal{B}(d_j, 0)$, $\mathcal{B}(d_j, 1)$ if $j \in \mathcal{D}_m$, and $1, \ldots, b_j$ otherwise. It is understood here that within this table, probabilities are conditioned on $B_j \geq d_j$ when B_j has logits of continuation type.

Example 3.1. Let B_1, B_2, B_3 have three categories each, $\mathcal{M}_1 = \{1, 2\}$ with logits l and g, $\mathcal{M}_2 = \{1, 3\}$ with logits l and c, and $\mathcal{M}_3 = \{1, 2, 3\}$ with logits l, g and c. Within $\mathcal{M}_1$, $\boldsymbol{d}_{\{2\}}$ is one-dimensional and can take value 1 or 2; each of these values corresponds to a 3×2 table where B_2 is dichotomized. The same holds for $\boldsymbol{d}_{\{3\}}$ within $\mathcal{M}_2$, except that when $\boldsymbol{d}_{\{3\}} = 2$, the table is conditioned to $B_3 > 1$. Within $\mathcal{M}_3$ there are four $3 \times 2 \times 2$ tables corresponding to the following values of $\boldsymbol{d}_{\{2,3\}}$: (1 1), (1 2), (2 1), (2 2); the second and forth are conditioned on $B_3 > 1$.

Since $\mathcal{D}_m$ is uniquely identified by the marginal $\mathcal{M}_m$, in the following we write $\boldsymbol{d}$ instead of $\boldsymbol{d}_{\mathcal{D}_m}$; however, when we need to refer to a subset of the elements of $\boldsymbol{d}$, the set of variables involved will be given explicitly. For a given $pdt(m, \boldsymbol{d})$, let $\mathcal{N}_m(\boldsymbol{d}) = \{j \in \mathcal{D}_m : d_j > 1\}$, $\mathcal{C}_m(\boldsymbol{d}) = \{j \in \mathcal{D}_m : d_j > 1, B_j \text{ of continuation type}\}$, and $\mathcal{A}_m(\boldsymbol{d}) = \{\mathcal{I} \in \mathcal{F}_m : \mathcal{N}_m(\boldsymbol{d}) \subseteq \mathcal{I}\}$. Note that $\mathcal{A}_m(\boldsymbol{d})$ is an *ascending class* of subsets of $\mathcal{F}_m$ while its complement $\mathcal{F}_m \backslash \mathcal{A}_m(\boldsymbol{d})$ is a *descending* class. Moreover, $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d}) = \mathcal{R}_m \cup (\mathcal{F}_m \backslash \mathcal{A}_m(\boldsymbol{d}))$ (where $\mathcal{R}_m = \mathcal{P}_m \backslash \mathcal{F}_m$), being the union of two descending and disjoint classes, is again a descending class of subsets of $\mathcal{P}_m$.

Example 3.2. Let B_1, B_2 and B_3 have three categories each, $\mathcal{M}_1 = \{1, 3\}$, $\mathcal{M}_2 = \{2, 3\}$, and suppose that within $\mathcal{M}_3 = \{1, 2, 3\}$, only B_1 has logits of local type; then $\mathcal{R}_3 = \{\{1\}, \{3\}, \{1, 3\}, \{2\}, \{2, 3\}\}$ and when $\boldsymbol{d}_{\{2,3\}} = (1 \quad 1)$, $\mathcal{A}_3(\boldsymbol{d}_{\{2,3\}}) = \mathcal{F}_3$ because $\mathcal{N}_3(\boldsymbol{d}_{\{2,3\}}) = \emptyset$. Instead, when $\boldsymbol{d}_{\{2,3\}} = (1 \quad 2)$, $\mathcal{N}_3(\boldsymbol{d}_{\{2,3\}}) = \{3\}$, $\mathcal{A}_3(\boldsymbol{d}_{\{2,3\}}) = \{\{1, 2, 3\}\}$ and $\mathcal{F}_3 \backslash \mathcal{A}_3(\boldsymbol{d}_{\{2,3\}}) = \{\{1, 2\}\}$. For a different example with all variables of non-local type, let $\mathcal{M}_1 = \{1, 3\}$ and $\mathcal{M}_2 = \{1, 2, 3\}$ so $\mathcal{R}_2 = \{\{1\}, \{3\}, \{1, 3\}\}$. When $\boldsymbol{d}_{\{1,2,3\}} = (1 \quad 1 \quad 1)$, $\mathcal{A}_2(\boldsymbol{d}_{\{1,2,3\}}) = \mathcal{F}_2 = \{\{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$ because $\mathcal{N}_2(\boldsymbol{d}_{\{1,2,3\}}) = \emptyset$; when $\boldsymbol{d}_{\{1,2,3\}} = (2 \quad 2 \quad 1)$, $\mathcal{N}_2(\boldsymbol{d}_{\{1,2,3\}}) = \{1, 2\}$, hence $\mathcal{A}_2(\boldsymbol{d}_{\{1,2,3\}}) = \{\{1, 2\}, \{1, 2, 3\}\}$ and $\mathcal{F}_2 \backslash \mathcal{A}_2(\boldsymbol{d}_{\{1,2,3\}}) = \{\{2\}, \{2, 3\}\}$.

When the same type of design matrix used to define the canonical parameters in the full joint distribution is applied to $pdt(m, \boldsymbol{d})$, the corresponding mean value parameters are the conditional probabilities

$$p\big(B_j > d_j,\ \forall j \in \mathcal{D}_m \cap \mathcal{I},\ \ B_j > x_j,\ \forall j \in \mathcal{I}\backslash\mathcal{D}_m \mid\ \ B_j > d_j - 1,\ \forall j \in \mathcal{C}_m(\boldsymbol{d})\big);$$

they are denoted by $\nu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}, \boldsymbol{x}_{\mathcal{I}\backslash\mathcal{D}_m}; \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})})$ and, by recalling the similar definition of the mean value parameters for the full joint distribution, it easily follows that we may write

$$\nu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}, \boldsymbol{x}_{\mathcal{I}\backslash\mathcal{D}_m}; \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})}) = \frac{\mu_{\mathcal{I}\cup\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}, \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})\backslash\mathcal{I}} - \mathbf{1}, \boldsymbol{x}_{\mathcal{I}\backslash\mathcal{D}_m})}{\mu_{\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})} - \mathbf{1})}, \tag{8}$$

where the denominator is constant within $pdt(m, \boldsymbol{d})$, and equal to 1 if $\mathcal{C}_m(\boldsymbol{d})$ is empty.

We now introduce suitable collections of mean value parameters and interaction parameters that play an important role below. Let $\mathcal{V}_m(\boldsymbol{d})$ denote the full collection of $\nu_{\mathcal{I}}$, $\mathcal{I} \in \mathcal{P}_m$, within $pdt(m, \boldsymbol{d})$; this can be partitioned into

$$\begin{aligned}
{}^{\uparrow}\mathcal{V}_m(\boldsymbol{d}) &= \{\nu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}, \boldsymbol{x}_{\mathcal{I}\backslash\mathcal{D}_m}; \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})}),\ \forall\, \mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})\}\\
{}^{\downarrow}\mathcal{V}_m(\boldsymbol{d}) &= \{\nu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}, \boldsymbol{x}_{\mathcal{I}\backslash\mathcal{D}_m}; \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})}),\ \forall\, \mathcal{I} \in \mathcal{P}_m\backslash\mathcal{A}_m(\boldsymbol{d})\},
\end{aligned}$$

where the notation is a reminder that $\mathcal{A}_m(\boldsymbol{d})$ is an ascending class and its complement with respect to $\mathcal{P}_m$ is descending. The following two sets of parameters may be uniquely associated with a given $pdt(m, \boldsymbol{d})$:

$$\begin{aligned}
{}^{\uparrow}\mathcal{E}_m(\boldsymbol{d}) &= \{\eta_{\mathcal{I};\mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}}) :\ \mathcal{I} \in \mathcal{F}_m,\ \boldsymbol{d}_{\mathcal{D}_m\backslash\mathcal{I}} = \mathbf{1},\ \boldsymbol{x}_{\mathcal{D}_m\cap\mathcal{I}} = \boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}\},\\
{}^{\uparrow}\mathcal{U}_m(\boldsymbol{d}) &= \{\mu_{\mathcal{I}}(\boldsymbol{x}_{\mathcal{I}}) :\ \mathcal{I} \in \mathcal{F}_m,\ \boldsymbol{d}_{\mathcal{D}_m\backslash\mathcal{I}} = \mathbf{1},\ \boldsymbol{x}_{\mathcal{D}_m\cap\mathcal{I}} = \boldsymbol{d}_{\mathcal{D}_m\cap\mathcal{I}}\}.
\end{aligned}$$

The previous sets provide a partition of the parameters $\eta_{\mathcal{I};\mathcal{M}}(\boldsymbol{x}_{\mathcal{I}})$ and $\mu_{\mathcal{I}}(\boldsymbol{x}_{\mathcal{I}})$ by assigning each of them to a $pdt(m, \boldsymbol{d})$ such that all the elements of $\mathcal{D}_m\backslash\mathcal{I}$ have cut point equal to 1, so that $\mathcal{N}_m(\boldsymbol{d}) \subseteq \mathcal{D}_m \cap \mathcal{I}$, implying $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$.

Example 3.3. Going back to Example 3.1, within $\mathcal{M}_1$, when $\boldsymbol{d}_{\{2\}} = 1$, ${}^{\uparrow}\mathcal{E}_1(d_{\{2\}})$ contains the two logits of B_1 (any cut point), the logit of B_2 and the log-odds ratios B_1, B_2 at cut point 1 for B_2. In the first part of Example 3.2, within $\mathcal{M}_3$, when $\boldsymbol{d}_{\{2,3\}} = (1\ 1)$, ${}^{\uparrow}\mathcal{E}_3(\boldsymbol{d}_{\{2,3\}})$ has elements $\eta_{\{1,2\},\{1,2,3\}}(1\ 1)$, $\eta_{\{1,2\},\{1,2,3\}}(2\ 1)$, $\eta_{\{1,2,3\},\{1,2,3\}}(1\ 1\ 1)$ and $\eta_{\{1,2,3\},\{1,2,3\}}(2\ 1\ 1)$. Instead, when $\boldsymbol{d}_{\{2,3\}} = (1\ 2)$, ${}^{\uparrow}\mathcal{E}_3(\boldsymbol{d}_{\{2,3\}})$ has only two elements: $\eta_{\{1,2,3\},\{1,2,3\}}(1\ 1\ 2)$ and $\eta_{\{1,2,3\},\{1,2,3\}}(2\ 1\ 2)$ because $\mathcal{A}_3(\boldsymbol{d}_{\{2,3\}})$ contains only the set $\{1, 2, 3\}$.

In the following assume that the *pdt*'s are ordered first with respect to $m = 1, \ldots, s$ and, for the same m, in lexicographic order of cut points; formally $pdt(h, \boldsymbol{c}) \prec pdt(m, \boldsymbol{d})$ if $h < m$ or $h = m$ and there exists an integer r such

that $c_r < d_r$, $c_j \leq d_j$, $\forall j < r$. Lastly, define

$$\mathcal{U}_m(\boldsymbol{d}) = \bigcup_{pdt(h,\boldsymbol{c}) \prec pdt(m,\boldsymbol{d})} {}^{\uparrow}\mathcal{U}_h(\boldsymbol{c}), \qquad \mathcal{E}_m(\boldsymbol{d}) = \bigcup_{pdt(h,\boldsymbol{c}) \prec pdt(m,\boldsymbol{d})} {}^{\uparrow}\mathcal{E}_h(\boldsymbol{c}),$$

$$\begin{aligned} {}^{\star}\mathcal{U}_m(\boldsymbol{d}) = \{\mu_{\mathcal{J}}(\boldsymbol{x}_{\mathcal{J}}) :\ & \mathcal{J} = \mathcal{I} \cup \mathcal{C}_m(\boldsymbol{d}),\ \mathcal{I} \in \mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d}),\ \boldsymbol{x}_{\mathcal{D}_m \cap \mathcal{I}} = \boldsymbol{d}_{\mathcal{D}_m \cap \mathcal{I}}, \\ & \boldsymbol{x}_{\mathcal{C}_m(\boldsymbol{d}) \backslash \mathcal{I}} = \boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d}) \backslash \mathcal{I}} - \boldsymbol{1},\ \boldsymbol{x}_{\mathcal{D}_m \backslash [\mathcal{I} \cup \mathcal{C}_m(\boldsymbol{d})]} = \boldsymbol{1}\}. \end{aligned}$$

Lemma 1. *For any $pdt(m,\boldsymbol{d}) \succ pdt(1,\boldsymbol{1})$, ${}^{\star}\mathcal{U}_m(\boldsymbol{d}) \subset \mathcal{U}_m(\boldsymbol{d})$, and the transformation from the elements of $\mathcal{U}_m(\boldsymbol{d})$ to the elements of $[\mathcal{U}_m(\boldsymbol{d}) \backslash {}^{\star}\mathcal{U}_m(\boldsymbol{d})] \cup {}^{\downarrow}\mathcal{V}_m(\boldsymbol{d})$ is a diffeomorphism.*

Proof of Lemma 1. As regards the first statement, for any given $\mu_{\mathcal{J}}(\boldsymbol{x}_{\mathcal{J}}) \in {}^{\star}\mathcal{U}_m(\boldsymbol{d})$ let $pdt(h,\boldsymbol{c})$ be the *pdt* such that $\mu_{\mathcal{J}}(\boldsymbol{x}_{\mathcal{J}}) \in {}^{\uparrow}\mathcal{U}_h(\boldsymbol{c})$. Since $\mathcal{J} = \mathcal{I} \cup \mathcal{C}_m(\boldsymbol{d}) \in \mathcal{P}_m$, we must have $h \leq m$. When $h = m$, by definition $c_j \leq d_j$ with $c_j < d_j$ for some j unless $d_j = 1$ for all $j \in \mathcal{D}_m \backslash \mathcal{I}$; but this is impossible because $\mathcal{I} \notin \mathcal{A}_m(\boldsymbol{d})$. Thus $pdt(h,\boldsymbol{c}) \prec pdt(m,\boldsymbol{d})$.

For the main part of the lemma, note that, when $\mathcal{I} \notin \mathcal{A}_m(\boldsymbol{d})$, (8) defines a transformation between the elements of ${}^{\star}\mathcal{U}_m(\boldsymbol{d})$ and those of ${}^{\downarrow}\mathcal{V}_m(\boldsymbol{d})$. The fact that the elements of ${}^{\downarrow}\mathcal{V}_m(\boldsymbol{d})$ are strictly less than one and distinct implies that $\mu_{\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})} - \boldsymbol{1})$, which appears in the denominator of (8), cannot be an element of ${}^{\star}\mathcal{U}_m(\boldsymbol{d})$, and that the elements of ${}^{\star}\mathcal{U}_m(\boldsymbol{d})$ are distinct. Thus the transformation is one-to-one.

The next lemma provides a *mixed parameterization* (Barndorff-Nielsen (1978, p.121)) for any *pdt*.

Lemma 2. *For any $pdt(m,\boldsymbol{d})$, the components of ${}^{\uparrow}\mathcal{E}_m(\boldsymbol{d})$ are variation independent from those of ${}^{\downarrow}\mathcal{V}_m(\boldsymbol{d})$, and the transformation from the elements of ${}^{\downarrow}\mathcal{V}_m(\boldsymbol{d}) \cup {}^{\uparrow}\mathcal{E}_m(\boldsymbol{d})$ to those of $\mathcal{V}_m(\boldsymbol{d})$ is a diffeomorphism.*

Proof of Lemma 2. We first show that the elements of ${}^{\uparrow}\mathcal{E}_m(\boldsymbol{d})$ are the interactions of local type defined on $pdt(m,\boldsymbol{d})$ for all $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$. Note that, while in the definition of $\eta_{\mathcal{I};\mathcal{M}_m}(\boldsymbol{x}_{\mathcal{I}})$ we have $B_j = 1, \forall j \in \mathcal{M}_m \backslash \mathcal{I}$, in the corresponding interaction parameters defined on the *pdt*, $B_j \in \mathcal{B}(x_j, 0) \neq 1$ whenever $j \in \mathcal{N}_m(\boldsymbol{d}) \backslash \mathcal{I}$; however this set is empty since $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$ implies $\mathcal{N}_m(\boldsymbol{d}) \subseteq \mathcal{I}$. The fact that all probabilities within the *pdt* are divided by $\mu_{\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})} - \boldsymbol{1})$ is irrelevant because the log of this probability appears an even number of times with opposite signs in the definition of the interaction and thus will cancel out.

The above shows that the elements of ${}^{\uparrow}\mathcal{E}_m(\boldsymbol{d})$ are log-linear contrasts of local type of the probabilities within the *pdt*, and thus are also equal to our definition of canonical parameters when applied to such a table. The result follows from the well-known property of the mixed parameterization of the exponential

family, where a canonical parameter is used when $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$, and a mean value parameter otherwise.

Lemma 3. *For any $pdt(m, \boldsymbol{d}) \succ pdt(1, \mathbf{1})$, the transformation from the elements of $\mathcal{U}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{V}_m(\boldsymbol{d})$ to the elements of $\mathcal{U}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{U}_m(\boldsymbol{d})$ is a diffeomorphism.*

Proof of Lemma 3. The elements of ${}^{\uparrow}\mathcal{V}_m(\boldsymbol{d})$ are the mean value parameters ν's for which $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$. This implies that $\mathcal{C}_m(\boldsymbol{d}) \backslash \mathcal{I}$ is empty because $\mathcal{N}_m(\boldsymbol{d}) \subseteq \mathcal{I}$ so that (8) may be written as $\nu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m \cap \mathcal{I}}, \boldsymbol{x}_{\mathcal{I} \backslash \mathcal{D}_m}) = \mu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{D}_m \cap \mathcal{I}}, \boldsymbol{x}_{\mathcal{I} \backslash \mathcal{D}_m}) / \mu_{\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})} - \mathbf{1})$. Thus, for all $\mathcal{I} \in \mathcal{A}_m(\boldsymbol{d})$ there is a one-to-one correspondence between the ν's and the μ's for any given value of the denominator $\mu_{\mathcal{C}_m(\boldsymbol{d})}(\boldsymbol{d}_{\mathcal{C}_m(\boldsymbol{d})} - \mathbf{1})$ which belongs to $\mathcal{U}_m(\boldsymbol{d})$.

3.2. Smoothness of the link function

We show below that $\boldsymbol{\eta}$ is a diffeomorphism of $\boldsymbol{\mu}$ and thus of $\boldsymbol{\lambda}$. We proceed by induction, by assuming that *pdt*'s are processed according to the total order defined above; this procedure resembles very closely the one used in BR.

Theorem 1. *For any complete and hierarchical parameterization defined as in* (6)*, the mapping between Ω and the space of log-linear parameters $\boldsymbol{\lambda}$ is a diffeomorphism.*

Proof of Theorem 1. When $m = 1$ and $\boldsymbol{d} = \mathbf{1}$, both $\mathcal{E}_1(\mathbf{1})$ and $\mathcal{U}_1(\mathbf{1})$ are empty. Because $\mathcal{P}_1 = \mathcal{A}_1(\mathbf{1})$, ${}^{\uparrow}\mathcal{E}_1(\mathbf{1})$ contains all log-linear parameters for $pdt(1, \mathbf{1})$. The basic properties of the exponential family imply that the mapping between the elements of ${}^{\uparrow}\mathcal{E}_1(\mathbf{1})$ and those of ${}^{\uparrow}\mathcal{U}_1(\mathbf{1}) = {}^{\uparrow}\mathcal{V}_1(\mathbf{1})$ is a diffeomorphism. Let $pdt(a, \boldsymbol{b})$ denote the first *pdt* which follows $pdt(1, \mathbf{1})$. By definition, ${}^{\uparrow}\mathcal{E}_1(\mathbf{1}) = \mathcal{E}_a(\boldsymbol{b})$ and ${}^{\uparrow}\mathcal{U}_1(\mathbf{1}) = \mathcal{U}_a(\boldsymbol{b})$; thus, the mapping between the elements of $\mathcal{E}_a(\boldsymbol{b})$ and those of $\mathcal{U}_a(\boldsymbol{b})$ is a diffeomorphism. Now let $pdt(s, \boldsymbol{z})$ denote the last *pdt* and suppose that, for a $pdt(m, \boldsymbol{d}) \prec pdt(s, \boldsymbol{z})$, we have proved that the mapping between the elements of $\mathcal{E}_m(\boldsymbol{d})$ and those of $\mathcal{U}_m(\boldsymbol{d})$ is a diffeomorphism. If we write this relationship more briefly as $\mathcal{E}_m(\boldsymbol{d}) \leftrightarrow \mathcal{U}_m(\boldsymbol{d})$, we have

$$
\begin{aligned}
\mathcal{E}_r(\boldsymbol{e}) &= [\mathcal{E}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{E}_m(\boldsymbol{d})] \leftrightarrow [\mathcal{U}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{E}_m(\boldsymbol{d})] \quad \text{(by assumption)} \\
&\leftrightarrow [\mathcal{U}_m(\boldsymbol{d}) \backslash^{\star} \mathcal{U}_m(\boldsymbol{d})] \cup^{\downarrow} \mathcal{V}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{E}_m(\boldsymbol{d}) \quad \text{(by Lemma 1)} \\
&\leftrightarrow [\mathcal{U}_m(\boldsymbol{d}) \backslash^{\star} \mathcal{U}_m(\boldsymbol{d})] \cup \mathcal{V}_m(\boldsymbol{d}) \quad \text{(by Lemma 2)} \\
&\leftrightarrow [\mathcal{U}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{U}_m(\boldsymbol{d})] = \mathcal{U}_r(\boldsymbol{e}) \quad \text{(by Lemmas 1 and 3).}
\end{aligned}
$$

Lastly, when $m = s$ and $\boldsymbol{d} = \boldsymbol{z}$, the inductive argument given above implies that the mapping between the elements of $[\mathcal{E}_s(\boldsymbol{z}) \cup^{\uparrow} \mathcal{E}_s(\boldsymbol{z})]$ and those of $[\mathcal{U}_s(\boldsymbol{z}) \cup^{\uparrow} \mathcal{U}_s(\boldsymbol{z})]$ is a diffeomorphism. Since the two sets contain, respectively, all the elements of $\boldsymbol{\eta}$ and all the elements of $\boldsymbol{\mu}$ for the full joint distribution of $B_1, \ldots, B_q$, the mapping between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, and therefore also that between $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$, is a diffeomorphism.

The following corollary of Theorem 1 will be useful in the next section.

Corollary 1. *For any complete and hierarchical parameterization defined as in* (6)*, the mapping between* Ω *and the space of log-linear parameters* $\boldsymbol{\lambda}$ *has continuous second derivatives at every* $\boldsymbol{\eta} \in \Omega$*.*

Proof of Corollary 1 Note that Π is open and that, from Theorem 1, it follows that Ω is open too. Since $\boldsymbol{\eta} = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{\pi})$ has continuous second derivatives, the corollary follows from Theorem 1 and from the Inverse Function Theorem (see, for example, Fleming (1977)).

3.3. Compatibility and variation independence

Within a given parameterization, the value of the joint probability vector $\boldsymbol{\pi}$ that corresponds to an assigned value of $\boldsymbol{\eta}$ may be reconstructed by a Newton type algorithm like the ones described by Glonek (1996) or by Colombi and Forcina (2001). However, in case of failure to converge, such algorithms cannot provide any information as to whether this is caused by numerical instability or whether $\boldsymbol{\eta} \notin \Omega$ and, when this is the case, which components of $\boldsymbol{\eta}$ are causing the problem. Instead, a reconstruction algorithm which follows the proof of Theorem 1 would stop at a given $pdt(m, \boldsymbol{d})$ because the mean values collected in the set $\mathcal{U}_m(\boldsymbol{d}) \cup^{\uparrow} \mathcal{U}_m(\boldsymbol{d})$ do not correspond to a vector of probabilities $\boldsymbol{\pi} \in \Pi$. We discuss the issue below in greater detail and propose an extension of the results obtained by BR to the case of generalized interactions; unfortunately, within this context, variation independence of the elements of $\boldsymbol{\eta}$ may arise only in very special circumstances.

We first state two intermediate results.

Lemma 4. *The class of the maximal elements of* $\mathcal{R}_m$*,* $m = 2, \ldots, s$*, is decomposable if and only if the sequence* $\mathcal{M}_1, \ldots, \mathcal{M}_s$ *is ordered decomposable.*

Proof of Lemma 4. Let $\mathcal{J}_m = \{j : \mathcal{M}_j \text{ is maximal within } \mathcal{M}_1, \ldots, \mathcal{M}_{m-1}\}$; the class of maximal elements of $\mathcal{R}_m$ is given by $\{\mathcal{M}_j \cap \mathcal{M}_m,\ j \in \mathcal{J}_m\}$ and is decomposable if and only if $\{\mathcal{M}_j,\ j \in \mathcal{J}_m\}$ is decomposable. Since this has to hold for all $m > 1$, the statement of the lemma follows.

Lemma 5. *If the class of the maximal elements of* $\mathcal{R}_m$ *is decomposable, a sufficient condition for the corresponding class of* $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d})$ *to be decomposable is that* $\mathcal{N}_m(\boldsymbol{d})$ *has at most two elements and is not contained in* $\bigcup_1^{m-1} \mathcal{M}_j$*.*

Proof of Lemma 5. If $\mathcal{N}_m(\boldsymbol{d}) = \{j\}$, the maximal elements of $\mathcal{R}_m$ are contained in $\mathcal{M}_m \backslash \{j\}$, the unique maximal element of $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d})$. If $\mathcal{N}_m(\boldsymbol{d}) = \{j_1, j_2\}$, the maximal elements of $\mathcal{R}_m$ are contained either in $\mathcal{M}_m \backslash \{j_1\}$ or $\mathcal{M}_m \backslash \{j_2\}$, the only two maximal elements of $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d})$; the class of these two subsets is decomposable.

Example 3.4. Let B_1, B_2 and B_3 have three categories, $\mathcal{M}_1 = \{1, 2\}$, $\mathcal{M}_2 = \{1, 2, 3\}$, and suppose that B_1 only has logits of local type within both marginals. Within $\mathcal{M}_2$, $\mathcal{R}_2 = \{\{1\}, \{2\}, \{1, 2\}\}$ and, with $\boldsymbol{d}_{\{2,3\}} = (2\ 2)$, $\mathcal{F}_2 \backslash \mathcal{A}_2(\boldsymbol{d}_{\{2,3\}}) = \{\{3\}, \{1, 3\}\}$. Consequently, the class of the maximal elements of $\mathcal{P}_2 \backslash \mathcal{A}_2(\boldsymbol{d}_{\{2,3\}})$, $\{\{1, 2\}, \{1, 3\}\}$, is decomposable.

Theorem 2. *A vector of generalized interaction parameters may be incompatible if, for a given $\mathcal{M}_m$,*

(i) *$m > 1$ and the class of maximal elements of $\mathcal{R}_m$ is not decomposable;*

(ii) *there are two vectors of cut points, $\boldsymbol{d}$ and $\boldsymbol{e}$, such that $d_j = e_j$ for $j \neq h$, $e_h = d_h + 1$ and B_h has logits of global type;*

(iii) *$m > 1$ and there exists a vector of cut points $\boldsymbol{d}$ such that the class of maximal elements of $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d})$ is not decomposable.*

Proof of Theorem 2. Case (i) follows from Theorem 4 in BR; it may also be derived from Lemma 1 by noting that, if the maximal elements of $\mathcal{R}_m$ form a non-decomposable class, the vector of ν's recovered from previous marginals may not be compatible. Case (ii) arises because with logits of global type, the constraints on the survival function $\mu_{\mathcal{I}}(\boldsymbol{d}_{\mathcal{I} \cap \mathcal{D}_m}, \boldsymbol{x}_{\mathcal{I} \backslash \mathcal{D}_m}) > \mu_{\mathcal{I}}(\boldsymbol{e}_{\mathcal{I} \cap \mathcal{D}_m}, \boldsymbol{x}_{\mathcal{I} \backslash \mathcal{D}_m})$ may be violated. Case (iii) follows from Lemma 5: if the class of maximal elements of $\mathcal{P}_m \backslash \mathcal{A}_m(\boldsymbol{d})$ is not decomposable, as in case (i), incompatible $\nu' s$ are possible.

Theorem 2 indicates that when logits of global type are used, the elements of $\boldsymbol{\eta}$ will never be variation independent. However we have

Corollary 2. *If the parameterization is ordered decomposable and within each marginal there are at most two variables with logits of continuation type that do not appear in previous marginals, and none with logits of global type, the elements of $\boldsymbol{\eta}$ are variation independent.*

Proof of Corollary 2. The result follows immediately from Theorem 2 and Lemma 4.

4. Likelihood Inference on Linear Equality and Inequality Constraints

We now provide an algorithm for maximum likelihood estimation and derive the asymptotic distribution of the likelihood ratio for testing linear equality and inequality constraints on $\boldsymbol{\eta}$ under multinomial sampling. Before going into the technical details, we motivate the class of models defined by such constraints with two examples. The novelty of the approach consists in combining familiar hypotheses, imposing the constraint that certain linear combinations of interaction parameters are equal to 0, with hypotheses stating that additional linear contrasts are non-negative. Constraints defined by linear inequalities may be

useful for stating that a given marginal (or conditional) distribution is stochastically larger than another, that two variables are positively dependent, or that the strength of the dependence increases with a third variable. The class of marginal models that can be defined by imposing a set of equality and inequality constraints on a link function, defined as in Glonek (1996), has been studied by Colombi and Forcina (2001). We briefly examine below a few additional models that are allowed within this extended formulation.

4.1. Examples of constrained models

Example 4.1. Suppose that B_1, B_2 and B_3 represent the *education* of father, mother and son, respectively, and that B_4 is the *social class of the son*. With $\mathcal{M}_1 = \{1\}$, $\mathcal{M}_2 = \{2\}$ and $\mathcal{M}_3 = \{1, 2\}$, we might consider several constraints, for example positive quadrant dependence (PQD), that B_1 is stochastically larger than B_2, or that the marginal logits are related by a constant shift (see Bartolucci, Forcina and Dardanoni (2001)). In $\mathcal{M}_4 = \{1, 2, 3\}$, we might investigate whether the conditional logits of global type for B_3 are increasing coordinate-wise in B_1 and B_2, or whether the two effects are additive. Lastly, in $\mathcal{M}_5 = \{1, 2, 3, 4\}$, we might examine whether B_4 is independent of B_1, B_2, given B_3.

Example 4.2. Suppose that *smoking habit* (B_1), *obesity* (B_2), *dyspnea* (B_3) and *heart murmur* (B_4) are ordinal categorical variables. First of all, we might formulate PQD within $\mathcal{M}_1 = \{1, 2\}$ and $\mathcal{M}_2 = \{1, 3\}$. Then, we might consider the marginal $\mathcal{M}_3 = \{1, 2, 4\}$ and, using global logits for B_4 and local logits for B_1 and B_2, formulate the hypothesis that the effects of B_1 and B_2 on B_4 are additive. The remaining interactions, defined within $\mathcal{M}_4 = \{1, 2, 3, 4\}$, might be used to formulate the hypothesis of conditional independence between B_2 and B_3, given the other two variables. If we were interested in modelling the effect of B_1 on B_2, we should insert the marginal $\{1\}$ before $\{1, 2\}$ with logits of local type for B_1 within $\{1, 2\}$; then the requirement that all the local-global log-odds ratios between B_1 and B_2 are non-negative implies an isotonic regression of B_2 on B_1. By inserting the marginal $\{1, 2, 3\}$ before $\{1, 2, 3, 4\}$, we can formulate, for example, the hypothesis of conditional independence between B_2 and B_3 given B_1, if this has logits of local type within $\{1, 2, 3\}$. All the remaining interaction parameters defined within $\{1, 2, 3, 4\}$ involve sets of variables that contain $\{3, 4\}$, thus they may be used to model the conditional distribution of B_3, B_4 given B_1, B_2, if the latter have logits of local type.

4.2. Formulation of constrained models

The class of constrained models studied in this section may be defined as

$$\mathcal{H}_{eu} = \{\boldsymbol{\eta} : \boldsymbol{E}\boldsymbol{\eta} = \mathbf{0}, \boldsymbol{U}\boldsymbol{\eta} \geq \mathbf{0}\},$$

where $\boldsymbol{E}$ is a full rank matrix with a rows; we also assume that $(\boldsymbol{E}' \quad \boldsymbol{U}')'$ has full rank b. Note that we are mainly interested in models that restrict the dependence structure of $B_1, \ldots, B_q$, and possibly in comparing different univariate marginal distributions having the same number of categories. In fact, we might be interested in testing for a stochastic ordering between two marginal distributions, or that one distribution is obtained by a constant shift on the logits of another marginal distribution. We will also consider the hypothesis $\mathcal{H}_{ee} = \{\boldsymbol{\eta} : \boldsymbol{E}\boldsymbol{\eta} = \boldsymbol{0}, \boldsymbol{U}\boldsymbol{\eta} = \boldsymbol{0}\}$ formulated by turning all the inequality constraints on $\boldsymbol{\eta}$ into equality constraints.

Remark 3. Since $\boldsymbol{\eta}$ would not be compatible unless global logits within the same marginal distribution are strictly decreasing, in order to ensure that $\mathcal{H}_{ee}$ is not empty, constraints that clash with this requirement cannot be allowed. This is not a limitation, however, because only hypotheses which are clearly incompatible are not allowed, for instance, two adjacent global logits being equal. On the other hand, contrasts aimed at comparing corresponding global logits of different marginals with the same number of categories are not affected.

4.3. Maximum likelihood estimation

In this section we describe an algorithm for computing maximum likelihood estimates of $\boldsymbol{\eta}$ under a suitable set of constraints. At each step of the algorithm, a quadratic approximation of the log-likelihood is maximized with respect to the vector $\boldsymbol{\lambda}$ of log-linear parameters under a linearized version of the constraints. This approach, which is related to that of Aitchison and Silvey (1958), avoids incompatibility problems that typically arise when we maximize the likelihood with respect to $\boldsymbol{\eta}$ directly, by means of the Newton-Raphson or the Fisher-scoring algorithms.

Consider first the case of no covariates, and let $\boldsymbol{y}$ be the vector of the frequencies observed in a sample of size n. Assuming multinomial sampling, the log-likelihood may be written as

$$L(\boldsymbol{\lambda}) = \boldsymbol{y}'\boldsymbol{G}\boldsymbol{\lambda} - n \log[\boldsymbol{1}' \exp(\boldsymbol{G}\boldsymbol{\lambda})] + \text{constant},$$

so that the score vector and the average information matrix are equal to $\boldsymbol{s}(\boldsymbol{\lambda}) = \boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})$ and $\boldsymbol{F}(\boldsymbol{\lambda}) = \boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G}$, respectively, where $\boldsymbol{\Omega}$=diag$(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$. Let $\boldsymbol{\lambda}^{(h)}$ denote the estimate after h steps; at step $h+1$ the algorithm at issue consists in maximizing a quadratic approximation of $L(\boldsymbol{\lambda})$, which has the same score vector and the same information matrix of $L(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^{(h)}$, under a linearized version of the constraints derived through the following first order approximation of $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \boldsymbol{\eta}^{(h)} + \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\lambda}'}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(h)}), \quad \text{where} \quad \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\lambda}'} = \boldsymbol{C}\,\text{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-1}\boldsymbol{M}\boldsymbol{\Omega}\boldsymbol{G}.$$

When there are r independent samples drawn from different strata, one for every configuration of the explanatory variables, the score function is obtained simply by stacking the vectors $\boldsymbol{s}(\boldsymbol{\lambda}_i)$, where $\boldsymbol{\lambda}_i$ denotes the vector of log-linear parameters for the i-th stratum. It can easily be shown that the expected information matrix is block diagonal with blocks $\boldsymbol{G}'\boldsymbol{\Omega}_i\boldsymbol{G}$, where $\boldsymbol{\Omega}_i$ denotes the variance kernel for the i-th stratum.

4.4. Asymptotic properties of the mle

Let $\dot{\boldsymbol{\eta}}$ be the unconstrained mle of $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$ be the mle under $\mathcal{H}_{eu}$ and $\bar{\boldsymbol{\eta}}$ be the mle under $\mathcal{H}_{ee}$. Provided $\boldsymbol{\eta}_0$, the true value of $\boldsymbol{\eta}$ under $\mathcal{H}_{ee}$, is an interior point of the parameter space, the three estimates $\dot{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{\eta}}$ exist and converges to $\boldsymbol{\eta}_0$ in probability. This is because our parameterization satisfies the two basic assumptions given by Rao (1973, Sec. 5e.2): 1.1, known as *strong identifiability*, and 2.1, continuity of the transformation from $\boldsymbol{\eta}$ to $\boldsymbol{\pi}$. Both assumptions follow easily from our Theorem 1. The same theorem implies that the average information matrix, $\boldsymbol{F}(\boldsymbol{\eta})$, is of full rank and may be computed as

$$\boldsymbol{F}(\boldsymbol{\eta}) = \boldsymbol{B}'\boldsymbol{F}(\boldsymbol{\lambda})\boldsymbol{B}, \quad \text{with} \quad \boldsymbol{B} = \frac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{\eta}'} = \left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\lambda}'}\right)^{-1}.$$

4.5. Asymptotic distribution of the likelihood ratio test

We derive the asymptotic distribution of the likelihood ratio (LR) for testing $\mathcal{H}_{eu}$ against the saturated model $\mathcal{S}$, $\Lambda_{eu} = 2[L(\dot{\boldsymbol{\eta}}) - L(\hat{\boldsymbol{\eta}})]$, and the LR for testing $\mathcal{H}_{ee}$ against $\mathcal{H}_{eu}$, $\Lambda_{ee} - \Lambda_{eu} = 2[L(\hat{\boldsymbol{\eta}}) - L(\bar{\boldsymbol{\eta}})]$. Let $\tilde{\boldsymbol{\eta}}$ be any of the maximum likelihood estimator of $\boldsymbol{\eta}_0$ defined at the beginning of the previous section, and let $\boldsymbol{H}$ denote the Hessian computed at $\boldsymbol{\eta}$, where $\mid \boldsymbol{\eta} - \boldsymbol{\eta}_0 \mid < \mid \tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \mid$. Note that from Corollary 1 it follows that the elements of $\boldsymbol{H}$ are continuous functions. Lastly, $\boldsymbol{s}_0$ and $\boldsymbol{F}_0$ denote the score function and the average information matrix computed at the true value $\boldsymbol{\eta}_0$ respectively. Consider a second-order Taylor-series expansion of the log-likelihood $L(\tilde{\boldsymbol{\eta}})$ around $\boldsymbol{\eta}_0$, replace $-\boldsymbol{H}/n$ with $\boldsymbol{F}_0$, add and subtract a quadratic form in $\boldsymbol{s}_0$ and let $\boldsymbol{\theta} = \sqrt{n}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ and $\boldsymbol{x} = \boldsymbol{F}_0^{-1}\boldsymbol{s}_0/\sqrt{n}$, so that we may write

$$\begin{aligned} L(\tilde{\boldsymbol{\eta}}) &= L(\boldsymbol{\eta}_0) + \boldsymbol{s}_0'(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)' - \frac{\sqrt{n}}{2}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)'\Big(-\frac{1}{n}\boldsymbol{H}\Big)\sqrt{n}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\ &= L(\boldsymbol{\eta}_0) + \frac{1}{2n}\boldsymbol{s}_0'\boldsymbol{F}_0^{-1}\boldsymbol{s}_0 - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{x})'\boldsymbol{F}_0(\boldsymbol{\theta} - \boldsymbol{x}) + o_p(|\boldsymbol{\theta}|^2). \end{aligned}$$

Since the maximum likelihood estimator $\tilde{\boldsymbol{\eta}}$ is consistent, Theorem 1 of Andrews (1999) implies that the last term in the previous expansion can be replaced by an

$o_p(1)$ term. Thus when $\boldsymbol{\eta}_0$ belongs to $\mathcal{H}_{ee}$, Λ_{eu} and $\Lambda_{eu} - \Lambda_{ee}$ are asymptotically equivalent to

$$Q_{eu} = \min_{\boldsymbol{\theta} \in \mathcal{H}_{eu}} (\boldsymbol{\theta} - \boldsymbol{x})' \boldsymbol{F}_0 (\boldsymbol{\theta} - \boldsymbol{x}), \tag{9}$$

$$Q_{ee} = \min_{\boldsymbol{\theta} \in \mathcal{H}_{ee}} (\boldsymbol{\theta} - \boldsymbol{x})' \boldsymbol{F}_0 (\boldsymbol{\theta} - \boldsymbol{x}) - Q_{eu} \tag{10}$$

respectively. According to the Central Limit Theorem, when $n \to \infty$, $\boldsymbol{x}$ converges in distribution to a normal random vector with density $N(\mathbf{0}, \boldsymbol{F}_0^{-1})$, so that the two quadratic forms in (9) and (10) converge to chi-bar-squared random variables (Shapiro (1988)), a mixture of chi-squared distributions. In particular, Q_{ee} and Q_{eu} are distributed as the squared norm of the projection of a normal random variable with density $N(\mathbf{0}, \boldsymbol{F}_0^{-1})$ onto the convex cone defined by $\mathcal{H}_{eu}$ and onto its dual, respectively. Both distributions depend on the same set of probability weights used in reverse order; these weights may be estimated by Monte Carlo simulation in order to achieve a given precision in the estimated p-values (see Dardanoni and Forcina (1998), Section 4.5, and Colombi and Forcina (2001), Section 5 for details). It must be recalled that, when inequality constraints are present, the familiar asymptotic chi-squared distribution must be replaced by the chi-bar-squared distribution, because the parametric space defined by $\mathcal{H}_{eu}$ is a convex cone and not an affine space.

For simplicity's sake, we have considered the case of no covariates; the extension to independent multinomials, one for each configuration of the explanatory variables, is straightforward and is not discussed in detail.

Appendix

Construction of the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$

The matrix $\boldsymbol{C}$ is block diagonal with blocks $\boldsymbol{C}_{\mathcal{I}}$, $\forall \mathcal{I} \in \mathcal{P}(\mathcal{Q})$, while $\boldsymbol{M}$ is obtained by stacking the matrices $\boldsymbol{M}_{\mathcal{I},\mathcal{M}}$. These components are defined as follows:

$$\boldsymbol{C}_{\mathcal{I}} = \bigotimes_{j=1}^{q} \boldsymbol{C}_{\mathcal{I},j}, \qquad \boldsymbol{C}_{\mathcal{I},j} = \begin{cases} (-\boldsymbol{I}_{b_j - 1} \quad \boldsymbol{I}_{b_j - 1}) & \text{if } j \in \mathcal{I} \\ 1 & \text{otherwise} \end{cases},$$

$$\boldsymbol{M}_{\mathcal{I},\mathcal{M}} = \bigotimes_{j=1}^{q} \boldsymbol{M}_{\mathcal{I},\mathcal{M},j}, \quad \boldsymbol{M}_{\mathcal{I},\mathcal{M},j} = \begin{cases} \begin{pmatrix} \boldsymbol{A}_{0,j}(v) \\ \boldsymbol{A}_{1,j}(v) \end{pmatrix} & \text{if } j \in \mathcal{I} \\ \boldsymbol{u}'_{b_j} & \text{if } j \in \mathcal{M} \backslash \mathcal{I} \\ \mathbf{1}'_{b_j} & \text{if } j \in \mathcal{Q} \backslash \mathcal{M}, \end{cases}$$

where $\boldsymbol{A}_{0,j}(v)'$ is an identity matrix for $v = l$ and c, and an upper triangular matrix for $v = g$ and r, without the last column, $\boldsymbol{A}_{1,j}(v)'$ is a $b_j \times b_j$ lower

triangular matrix of ones for $v = g$ and c, and an identity matrix of the same size for $v = l$ and r without the first column. The vector $\boldsymbol{u}'_{b_j}$ is a row vector with the first element equal to one and the others $b_j - 1$ equal to zero, while $\mathbf{1}_{b_j}$ is a vector of ones of length b_j.

Left inverse of the matrix $\boldsymbol{G}$

By exploiting the properties of the Kronecker product, it is easily verified that $\boldsymbol{K}$ must have blocks

$$\boldsymbol{K}_{\mathcal{I}} = \bigotimes_{j=1}^{q} \boldsymbol{K}_{\mathcal{I},j}, \text{ where } \boldsymbol{K}_{\mathcal{I},j} = \begin{cases} \boldsymbol{D}_j & \text{if } j \in \mathcal{I} \\ \boldsymbol{u}_{b_j} & \text{otherwise} \end{cases},$$

and where $\boldsymbol{D}_j$ is the $(b_j - 1) \times b_j$ matrix of differences between adjacent terms. For example, with $\mathrm{Q} = \{1, 2, 3\}$, $\boldsymbol{K}_1 \log(\boldsymbol{\pi})$ are the adjacent logits of $\{1\}$ with $\{2, 3\}$ set at their initial levels. Note also that $\boldsymbol{T}_j$ is the right inverse of $\boldsymbol{D}_j$, that is $\boldsymbol{D}_j \boldsymbol{T}_j = \boldsymbol{I}$. Now let $\boldsymbol{\mu} = \boldsymbol{G}'\boldsymbol{\pi}$ denote the vector of mean value parameters corresponding to $\boldsymbol{\lambda}$. Since $\boldsymbol{T}'_j \boldsymbol{q}$ is the survival function of a vector $\boldsymbol{q}$ of b_j probabilities, it follows that the element $\mu_{\mathcal{I}}(\boldsymbol{x}_{\mathcal{I}})$ of $\boldsymbol{\mu}$ is $p(B_j > x_j, \forall j \in \mathcal{I})$.

Acknowledgements

We are very grateful to two referees for stimulating suggestions. We acknowledge the financial support of a MIUR 2002 grant.

References

Aitchison, J. and Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813-828.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67**, 1341-1383.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

Bartolucci, F. and Forcina, A. (2002). Extended RC Association Models allowing for Order Restrictions and Marginal Modelling. *J. Amer. Statist. Assoc.* **97**, 1192-1199.

Bartolucci, F., Forcina, A. and Dardanoni, V. (2001). Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. *J. Amer. Statist. Assoc.* **96**, 1497-1505.

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

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika* **88**, 1007-1019.

Dardanoni, V. and Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Amer. Statist. Assoc.* **93**, 1112-1123.

Douglas, R., Fienberg, S. E., Lee M. T., Sampson, A. R. and Whitaker, L. R. (1990). Positive dependence concepts for ordinal contingency tables. In *Topics in Statistical Dependence*, Institute of Mathematical Statistics Lecture Notes (Edited by H. W. Block, A. R. Sampson and T. H. Sanits). Haywar, California.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141-151.

Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. (1993). Regression models for discrete longitudinal responses. *Statist. Sci.* **8**, 284-309.

Fleming, W. (1977). *Functions of Several Variables*. Springer, Berlin.

Glonek, G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika* **83**, 15-28.

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

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

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

Lang, J. B. and Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**, 626-632.

Lauritzen, S. (1996). *Graphical Models*. Oxford Science Publications, Oxford.

Molenberghs, G. and Lesaffre, E. (1994). Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *J. Amer. Statist. Assoc.* **89**, 633-644.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd edition. Wiley, New York.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate Analysis. *Internat. Statist. Rev.* **56**, 49-62.

Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli 20, 06123 Perugia, Italy.

E-mail: bart@stat.unipg.it

Dipartimento di Ingegneria gestionale e dell'Informazione, Viale Marconi, 5, 24044 Dalmine, Italy.

E-mail: roberto.colombi@unibg.it

Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli 20, 06123 Perugia, Italy.

E-mail: forcina@stat.unipg.it

(Received March 2005; accepted October 2005)