MARGINAL MODELS FOR CATEGORICAL DATA

By Wicher P. Bergsma[1] and Tamás Rudas[2]

Tilburg University and Eötvös Loránd University

Statistical models defined by imposing restrictions on marginal distributions of contingency tables have received considerable attention recently. This paper introduces a general definition of marginal log-linear parameters and describes conditions for a marginal log-linear parameter to be a smooth parameterization of the distribution, and to be variation independent. Statistical models defined by imposing affine restrictions on the marginal log-linear parameters are investigated. These models generalize ordinary log-linear and multivariate logistic models. Sufficient conditions for a log-affine marginal model to be nonempty, and to be a curved exponential family are given. Standard large sample theory is shown to apply to maximum likelihood estimation of log-affine marginal models for a variety of sampling procedures.

1. Introduction. Several recent papers discuss the theory and application of models for contingency tables which impose restrictions on marginal distributions of the contingency table, see, for example, McCullagh and Nelder (1989), Liang, Zeger, and Qaqish (1992), Becker (1994), Lang and Agresti (1994), Glonek and McCullagh (1995), and Bergsma (1997). While these models are flexible and useful, certain theoretical questions have remained open in the literature. These include, firstly, the existence of a joint distribution with certain restrictions on some of its marginals, or general conditions under which the existence of such distributions is guaranteed; secondly, the determination of the dimension of a model; thirdly, conditions for the applicability of large sample results for maximum likelihood estimates.

To illustrate the importance of the above questions, consider a $2 \times 2 \times 2$ contingency table ABC, and the AB, BC, and AC marginal tables. Assume that in the first two marginal tables the cells $(1,1)$ and $(2,2)$, while in the last table the cells $(1,2)$ and $(2,1)$ have probabilities equal to $1/2$. Although these marginals are (weakly) compatible, because they imply uniform one-way marginal distributions, there exists no three-way distribution with these two-way marginals.

The concept of variation independence of parameters plays an important role in answering the questions above. Two parameters are variation independent when the range of possible values of one of them does not depend on the other's value.

---

[1]Bergsma's research was supported by The Netherlands Organization for Scientific Research (NWO), Project Number 400-20-001.

[2]Rudas's research was suppported by Grants T-016032 and T-032213 from the Hungarian National Science Foundation.

Both authors did part of their research while working at the Center for Applied Statistics, Central European University, in Budapest.

AMS 1991 subject classifications. Primary 62H17; secondary 62E99

Key words and phrases. marginal log-linear parameters, log-affine and log-linear marginal models, smooth parameterization, variation independence, existence and connectedness of a model, curved exponential family, asymptotic normality of maximum likelihood estimates

When two parameters are not variation independent, this causes problems in their interpretation, it leads to the possibility of the definition of non-existing models, and frequently also causes problems in various computations. Note that the multivariate logistic parameters of Glonek and McCullagh (1995) are not variation independent if there are more than two variables.

In the present paper, the above questions will be discussed under a general definition of marginal log-linear parameters. These are log-linear parameters computed from marginals of the joint distribution. Therefore, a marginal log-linear parameter is characterized by two subsets of the variables, one defining the marginal to which it pertains, and the other one defining the subset of this marginal for which it is computed.

Section 2 of the paper considers classes consisting of such ordered pairs of subsets. In this section, certain combinatorial properties of such classes, including ordered decomposability are defined. This property, for the models discussed in this paper, plays a role similar to that of decomposability in the case of ordinary log-linear models.

In Section 3, marginal log-linear parameters are defined, which are a generalization of both the ordinary log-linear parameters and of the parameters obtained by the multivariate logistic transform of Glonek and McCullagh (1995). In fact, these two parameterizations represent the two endpoints of a wide spectrum of possible marginal log-linear parameterizations. The main results of the section include that for classes of ordered subsets with a certain hierarchy property, parameterizations based on marginal log-linear parameters are smooth. For such smooth parameterizations, a necessary and sufficient condition for variation independence of the coordinates is that the marginals involved form an ordered decomposable set. In the latter case, repeated application of the iterative proportional fitting procedure can always be used to reconstruct the joint distribution from the values of marginal log-linear parameters.

In Section 4, log-affine marginal models are defined by restricting the values of certain marginal log-linear parameters. This class of models generalizes ordinary log-linear and multivariate logistic models and contains models which do not seem to have been considered before. The main result of the section establishes that a hierarchical marginal model is a curved exponential family, and it is proved that log-affine marginal models, if based on an ordered decomposable class of marginal log-linear parameters, are not empty. As an application, it is shown that log-affine marginal models can also be used to describe many types of sampling procedures, which provides a unified view of sampling and model restrictions. Also, Whittemore's (1978) collapsibility conditions can be described by a log-linear marginal model.

Finally, in Section 5, standard large sample theory is shown to apply to certain log-affine marginal models, implying asymptotic normality of the maximum likelihood estimates. The large sample results hold under a wide range of sampling procedures, such as Poisson, multinomial, and other distributions with given marginals. This section, however, does not attempt to discuss maximum likelihood estimation of log-affine marginal models in any depth, and detailed results concerning existence and uniqueness of maximum likelihood estimates will be considered in a forthcoming manuscript.

2. Sets of ordered pairs of subsets Let $\mathcal{V}$ be a finite set. The pair $(\mathcal{L},\mathcal{M})$, with $\mathcal{L} \subseteq \mathcal{M} \subseteq \mathcal{V}$, is an ordered pair of subsets of $\mathcal{V}$, where the ordering is with respect to inclusion. For a set $\mathcal{P}$ of ordered pairs, let

$$\mathbb{M} = \{\mathcal{M} \mid \exists \mathcal{L} \subseteq \mathcal{V} : (\mathcal{L},\mathcal{M}) \in \mathcal{P}\}.$$

For a certain ordering $\mathcal{M}_1,\ldots,\mathcal{M}_s$ of the elements of $\mathbb{M}$ and for $i = 1,\ldots,s$, let

$$\mathbb{L}_i = \{\mathcal{L} \mid (\mathcal{L},\mathcal{M}_i) \in \mathcal{P}\} \tag{2.1}$$

and let

$$\mathbb{K}_1 = \mathbb{P}(\mathcal{M}_1) \tag{2.2}$$

and for $i = 2,\ldots,s$ let

$$\mathbb{K}_i = \mathbb{P}(\mathcal{M}_i) \setminus (\mathbb{P}(\mathcal{M}_1) \cup \ldots \cup \mathbb{P}(\mathcal{M}_{i-1})) \tag{2.3}$$

where $\mathbb{P}(\mathcal{V})$ is the class of subsets of $\mathcal{V}$. Note that $\mathbb{L}_i$ is not empty and, for $i \neq j$,

$$\mathbb{K}_i \cap \mathbb{K}_j = \varnothing. \tag{2.4}$$

To illustrate, suppose $\mathcal{V} = \{A, B, C\}$ and

$$\mathcal{P} = \{(\{A\},\{A,B\}),(\{B\},\{A,B\}),(\{A,C\},\{A,B,C\})\} \tag{2.5}$$

Then $\mathbb{M} = \{\{A,B\},\{A,B,C\}\}$, $\mathbb{L}_1 = \{\{A\},\{B\}\}$, $\mathbb{L}_2 = \{\{A,C\}\}$, $\mathbb{K}_1 = \{\varnothing,\{A\},\{B\},\{A,B\}\}$, and $\mathbb{K}_2 = \{\{C\},\{A,C\},\{B,C\},\{A,B,C\}\}$.

Below, three properties applicable to sets of ordered pairs of subsets $\mathcal{P}$ are defined: hierarchy, completeness, and ordered decomposability.

The set $\mathcal{P}$ is called *hierarchical* if it has an ordering $\mathcal{M}_1,\ldots,\mathcal{M}_s$ of the elements of $\mathbb{M}$ such that

$$\mathcal{M}_i \not\subseteq \mathcal{M}_j \qquad \text{if } i > j \tag{2.6}$$

$$\mathbb{L}_i \subseteq \mathbb{K}_i. \tag{2.7}$$

Then $\mathcal{M}_1,\ldots,\mathcal{M}_s$ is called a hierarchical ordering of the elements of $\mathbb{M}$. Note that the $\mathcal{P}$ defined in (2.5) is hierarchical. Some examples of non-hierarchical sets $\mathcal{P}$ are given next.

EXAMPLE 1. *Suppose $\mathcal{V} = \{A,B\}$. If $\mathcal{P} = \{(\{A\},\{A\}), (\{A\},\mathcal{V})\}$ then $\mathbb{L}_1 = \mathbb{L}_2 = \{\{A\}\}$. Hence, if (2.7) holds, (2.4) cannot hold, therefore $\mathcal{P}$ is non-hierarchical. If $\mathcal{P} = \{(\{A\},\{A\}), (\varnothing,\mathcal{V})\}$, then $\mathbb{M} = \{\{A\},\mathcal{V}\}$. For (2.6) to hold, $\mathcal{M}_1 = \{A\}$ and $\mathcal{M}_2 = \mathcal{V}$, so that*

$$\mathbb{K}_1 = \{\varnothing,\{A\}\} \qquad \mathbb{L}_1 = \{\{A\}\}$$

$$\mathbb{K}_2 = (\mathbb{P}(\mathcal{V}) \setminus \mathbb{K}_1) = \{\{B\},\{A,B\}\} \qquad \mathbb{L}_2 = \{\varnothing\}$$

Since $\mathbb{L}_2 \not\subseteq \mathbb{K}_2$, (2.7) is violated, so $\mathcal{P}$ is non-hierarchical.

The set $\mathcal{P}$ is called *complete* if, for all $\mathcal{L} \subseteq \mathcal{V}$, there is exactly one $\mathcal{M} \subseteq \mathcal{V}$ such that $(\mathcal{L}, \mathcal{M}) \in \mathcal{P}$. Note that the $\mathcal{P}$ defined in (2.5) is incomplete. It follows that $\mathcal{P}$ is hierarchical *and* complete if and only if there is a hierarchical ordering of the elements of $\mathbb{M}$ for which $\mathcal{M}_s = \mathcal{V}$ and there is equality for all i in (2.7). A list of all hierarchical and complete $\mathcal{P}$s for which $|\mathcal{V}| \leq 2$ is given below.

Example 2. *If $\mathcal{V} = \{A\}$, there are two possible hierarchical and complete $\mathcal{P}$s:*

$$\begin{array}{l} \{(\varnothing,\varnothing), \quad (\{A\},\{A\})\} \\ \{(\varnothing,\{A\}), (\{A\},\{A\})\} \end{array}$$

and if $\mathcal{V} = \{A, B\}$, there are nine different possibilities:

$$\begin{array}{lllll}
\{\ (\varnothing,\varnothing), & (\{A\},\{A\}), & (\{B\},\{B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\{A\}), & (\{A\},\{A\}), & (\{B\},\{B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\{B\}), & (\{A\},\{A\}), & (\{B\},\{B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\varnothing), & (\{A\},\{A,B\}), & (\{B\},\{B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\{B\}), & (\{A\},\{A,B\}), & (\{B\},\{B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\varnothing), & (\{A\},\{A\}), & (\{B\},\{A,B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\{A\}), & (\{A\},\{A\}), & (\{B\},\{A,B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\varnothing), & (\{A\},\{A,B\}), & (\{B\},\{A,B\}), & (\{A,B\},\{A,B\})\ \} \\
\{\ (\varnothing,\{A,B\}), & (\{A\},\{A,B\}), & (\{B\},\{A,B\}), & (\{A,B\},\{A,B\})\ \}.
\end{array}$$

The number of different hierarchical and complete sets $\mathcal{P}$ increases faster than exponentially with the number of elements of $\mathcal{V}$.

A partial order for complete sets $\mathcal{P}$ is defined as follows. If $\mathcal{P} = \{(\mathcal{L}_i, \mathcal{M}_i) | \mathcal{L}_i \subseteq \mathcal{V}\}$ and $\mathcal{P}' = \{(\mathcal{L}_i, \mathcal{M}'_i) | \mathcal{L}_i \subseteq \mathcal{V}\}$ are complete, then $\mathcal{P} \ll \mathcal{P}'$ if $\mathcal{M}_i \subseteq \mathcal{M}'_i$ for all i. In this partial order,

$$\mathcal{P}_{\min} = \{(\mathcal{L}, \mathcal{L}) \mid \mathcal{L} \subseteq \mathcal{V}\}$$

is uniquely minimal, and

$$\mathcal{P}_{\max} = \{(\mathcal{L}, \mathcal{V}) \mid \mathcal{L} \subseteq \mathcal{V}\}$$

is uniquely maximal. For example, if $\mathcal{V} = \{A, B\}$, we find

$$\mathcal{P}_{\min} = \{\ (\varnothing,\varnothing)\ ,\ (\{A\},\{A\})\ ,\ (\{B\},\{B\})\ ,\ (\mathcal{V},\mathcal{V})\}$$
$$\mathcal{P}_{\max} = \{\ (\varnothing,\mathcal{V})\ ,\ (\{A\},\mathcal{V})\ ,\ (\{B\},\mathcal{V})\ ,\ (\mathcal{V},\mathcal{V})\ \}.$$

It is easy to verify that, for any $\mathcal{V}$, both $\mathcal{P}_{\min}$ and $\mathcal{P}_{\max}$ are hierarchical.

A class of incomparable (with respect to inclusion) subsets of $\mathcal{V}$ is called *decomposable* if it has at most two elements or if there is an ordering $\mathcal{M}_1, \ldots, \mathcal{M}_s$ of its elements such that, for $k = 3, \ldots, s$, there exists a $j_k < k$ such that

$$\left(\cup_{i=1}^{k-1} \mathcal{M}_i\right) \cap \mathcal{M}_k = \mathcal{M}_{j_k} \cap \mathcal{M}_k$$

(Haberman, 1974, page 181). A class of arbitrary subsets of $\mathcal{V}$ is *ordered decomposable* if it has at most two elements or if there is an ordering $\mathcal{M}_1, \ldots, \mathcal{M}_s$ of its elements such that (2.6) holds, and, for $k = 3, \ldots, s$, the maximal elements of $\{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$ form a decomposable set. The ordering $\mathcal{M}_1, \ldots, \mathcal{M}_s$ is then also called ordered decomposable. A set $\mathcal{P}$ is ordered decomposable if the elements of $\mathbb{M}$ have an ordering which is both hierarchical and ordered decomposable. Note that ordered decomposability is a generalization of decomposability which also applies when subsets are comparable. For incomparable subsets, the two concepts are the same.

Example 3. *If $\mathcal{P} = \mathcal{P}_{\max}$, $\mathbb{M} = \{\mathcal{V}\}$, so $\mathcal{P}_{\max}$ is ordered decomposable. On the other hand, if $\mathcal{P} = \mathcal{P}_{\min}$, $\mathbb{M} = \mathbb{P}(\mathcal{V})$, so $\mathcal{P}$ is not ordered decomposable unless $|\mathcal{V}| \leq 2$. For instance, if $\mathcal{V} = \{A, B, C\}$, then*

$$\mathbb{M} = \{\varnothing, \{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{A, C\}, \{A, B, C\}\}.$$

In a hierarchical ordering, $\mathcal{M}_8 = \{A, B, C\}$. But then the set of maximal elements of $\{\mathcal{M}_1, \dots, \mathcal{M}_7\}$ is $\{\{A, B\}, \{B, C\}, \{A, C\}\}$, which is not decomposable. Hence $\mathcal{P}$ is not ordered decomposable. Ordered decomposability can be obtained by, for example, replacing $(\{A, B\}, \{A, B\})$ in $\mathcal{P}$ by $(\{A, B\}, \{A, B, C\})$.

The remaining part of this section demonstrates that for any incomplete hierarchical $\mathcal{P}$ a complete and hierarchical $\overline{\mathcal{P}} \supset \mathcal{P}$ can be constructed which retains the fundamental properties of $\mathcal{P}$. For given hierarchical $\mathcal{P}$, let $\mathcal{M}_1, \dots, \mathcal{M}_s$ be a hierarchical ordering of the elements of $\mathbb{M}$. Then $\overline{\mathcal{P}}$ is defined as

$$\overline{\mathcal{P}} = \{(\mathcal{L}, \mathcal{M}_i) | \mathcal{L} \in \mathbb{K}_i, i \leq s\} \cup \{(\mathcal{L}, \mathcal{V}) | \mathcal{L} \not\subseteq \mathcal{M}_i \forall i \leq s\}.$$

For example, if $\mathcal{P}$ is defined as in (2.5), then

$$\begin{aligned}\overline{\mathcal{P}} = \{&(\varnothing, \{A, B\}), \{A\}, \{A, B\}), (\{B\}, \{A, B\}), (\{A, B\}, \{A, B\})\} \\ &\cup \{(\{C\}, \{A, B, C\}), (\{A, C\}, \{A, B, C\}), (\{B, C\}, \{A, B, C\}), (\{A, B, C\}, \{A, B, C\})\}.\end{aligned}$$

Note that the set $\overline{\mathcal{P}}$ depends only on $\mathcal{V}$ and $\mathbb{M}$, including its ordering. If $\overline{\mathbb{M}}$, $\overline{\mathbb{K}}_i$ and $\overline{\mathbb{L}}_i$ are defined similarly for $\overline{\mathcal{P}}$ as $\mathbb{M}$, $\mathbb{K}_i$ and $\mathbb{L}_i$ are defined for $\mathcal{P}$, it follows that

$$\overline{\mathbb{M}} = \mathbb{M} \cup \{\mathcal{V}\} \tag{2.8}$$

$$\overline{\mathbb{L}}_i = \overline{\mathbb{K}}_i \tag{2.9}$$

for $i \leq |\overline{\mathbb{M}}|$. Different hierarchical orderings of the elements of $\mathbb{M}$ may yield different $\overline{\mathcal{P}}$s. Whenever we write $\overline{\mathcal{P}}$, we refer to a class yielded by some ordering of the elements of $\mathbb{M}$. We have

Lemma 1.
A. $\mathcal{P} \subseteq \overline{\mathcal{P}}$
B. $\overline{\mathcal{P}}$ is hierarchical and complete
C. $\overline{\mathcal{P}}$ is ordered decomposable if and only if $\mathcal{P}$ is ordered decomposable

Proof. Part A follows from (2.7). Part B follows from (2.8) and (2.9). Part C is directly implied by (2.8).

3. Marginal log-linear parameters In Section 3.1 marginal frequencies are considered, and in Section 3.2 the definition of marginal log-linear parameters is given. In Section 3.3, smoothness properties of parameterizations of distributions on contingency tables in terms of marginal log-linear parameters are derived. In Section 3.4, necessary and sufficient conditions for marginal log-linear parameters to be variation independent are provided.

3.1. *Marginal frequencies* Let $\mathcal{V} = \{V_1, \dots V_p\}$ be a set of categorical variables, with V_j taking on values in the nonempty finite set $\mathcal{I}_j$, $1 \leq j \leq p$. The Cartesian product $\mathcal{T} = \times_{j=1}^{p} \mathcal{I}_j$ is a contingency table, with $\mathbf{i} = (i_1, \dots, i_p)$, for $i_j \in \mathcal{I}_j$, being a cell of the table. A non-negative real number $\mu(\mathbf{i})$ is called a cell frequency belonging to cell $\mathbf{i}$.

Let $\mathcal{F}$ be the class of strictly positive frequency distributions μ on $\mathcal{T}$. A function $\theta : \mathcal{F} \to \mathbf{R}^k$ $(k \geq 1)$ is called a *parameter* of $\mathcal{F}$.

If $\mathcal{M} \subseteq \mathcal{V}$ is a subset of the variables, then $\mathbf{i}_{\mathcal{M}}$ denotes a vector of those indices from $\mathbf{i}$, which belong to the variables in $\mathcal{M}$, that is, $(.)_{\mathcal{M}}$ is a projection operator. The collection of these marginal cells is the marginal table $\mathcal{T}_{\mathcal{M}}$. A *marginal frequency* $\mu_{\mathcal{M}}(\mathbf{i}_{\mathcal{M}})$ is obtained by appropriate summation of the cell frequencies $\mu(\mathbf{i})$. That is, a marginal frequency pertaining to marginal table $\mathcal{T}_{\mathcal{M}}$ is defined as follows:

$$\mu_{\mathcal{M}}(\mathbf{i}_{\mathcal{M}}) = \sum_{\mathbf{j} \in \mathcal{T} : \mathbf{j}_{\mathcal{M}} = \mathbf{i}_{\mathcal{M}}} \mu(\mathbf{j}).$$

The marginals $\mu^{(1)}_{\mathcal{M}_1}, \dots, \mu^{(s)}_{\mathcal{M}_s}$ $(\mathcal{M}_i \subseteq \mathcal{V})$ are said to be *weakly compatible* if

$$(\mu^{(i)}_{\mathcal{M}_i})_{\mathcal{M}_i \cap \mathcal{M}_j} = (\mu^{(j)}_{\mathcal{M}_j})_{\mathcal{M}_i \cap \mathcal{M}_j}$$

for $1 \leq i \leq j \leq s$. The marginals $\mu^{(1)}_{\mathcal{M}_1}, \dots, \mu^{(s)}_{\mathcal{M}_s}$ with a certain prescribed value are said to be *strongly compatible* if there exists a joint distribution μ such that $\mu^{(i)}_{\mathcal{M}_i} = (\mu)_{\mathcal{M}_i}$ for $1 \leq i \leq s$. Theorem 1 gives the necessary and sufficient condition for the former to imply the latter. It is a well-known result, and one direction of the proof follows from a counterexample (similar to the example of the three marginal tables with prescribed values in Section 1), while the other direction follows from the construction of Darroch, Lauritzen, and Speed (1980).

Theorem 1. *For a class of incomparable subsets $\{\mathcal{M}_1, \dots, \mathcal{M}_s\} \subseteq \mathbb{P}(\mathcal{V})$, weak compatibility of the marginal frequencies $\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_s}$ implies strong compatibility if and only if $\{\mathcal{M}_1, \dots, \mathcal{M}_s\}$ is decomposable.*

3.2. *Marginal log-linear parameters* By analogy to ordinary log-linear parameters for $\mathbf{i} \in \mathcal{T}$, marginal log-linear parameters pertaining to a marginal $\mathcal{M}$ are defined in the following recursive way:

$$\lambda^{\mathcal{M}}_{\varnothing}(\mathbf{i}_{\varnothing}) = \frac{1}{|\mathcal{T}_{\mathcal{M}}|} \sum_{\mathbf{j} \in \mathcal{T}_{\mathcal{M}}} \log \mu_{\mathcal{M}}(\mathbf{j}) \tag{3.10}$$

$$\lambda^{\mathcal{M}}_{\mathcal{L}}(\mathbf{i}_{\mathcal{L}}) = \frac{1}{|\mathcal{T}_{\mathcal{M} \setminus \mathcal{L}}|} \sum_{\mathbf{j} \in \mathcal{T}_{\mathcal{M}} : \mathbf{j}_{\mathcal{L}} = \mathbf{i}_{\mathcal{L}}} \log \mu_{\mathcal{M}}(\mathbf{j}) - \sum_{\mathcal{L}' \subset \mathcal{L}} \lambda^{\mathcal{M}}_{\mathcal{L}'}(\mathbf{i}_{\mathcal{L}'}). \tag{3.11}$$

The (first) term on the right hand side of these equations corresponds to an average over those marginal cells in table $\mathcal{T}_{\mathcal{M}}$ whose $\mathcal{L}$-index is the same as that of $\mathbf{i}_{\mathcal{M}}$. Note that the parameters $\lambda_{\mathcal{L}}^{\mathcal{V}}$ are the ordinary log-linear parameters.

There are other ways than (3.10) and (3.11) to define the marginal log-linear parameters. Our definition corresponds to so-called effect-coding. Glonek and McCullagh (1995) used so-called dummy coding, which yields different but equivalent parameters; without going into the technical details here, the results of this article hold for both types of parameters.

A set of marginal log-linear parameters is characterized by a set of ordered pairs. For a set of ordered pairs $\mathcal{P}$, let

$$\lambda_{\mathcal{P}} = \left\{\lambda_{\mathcal{L}}^{\mathcal{M}} : (\mathcal{L}, \mathcal{M}) \in \mathcal{P}\right\}.$$

The parameter $\lambda_{\mathcal{P}}$ with a certain prescribed value is said to be *strongly compatible* if there exists a μ which yields $\lambda_{\mathcal{P}}$.

The ordinary log-linear parameters are $\lambda_{\mathcal{P}_{\max}}$, and the multivariate logistic transform parameters (McCullagh & Nelder, 1989; Glonek & McCullagh, 1995) are $\lambda_{\mathcal{P}_{\min}}$. If $\mathcal{V} = \{A, B, C\}$, one obtains

$$\lambda_{\mathcal{P}_{\max}} = \left\{\lambda_{\varnothing}^{ABC}, \lambda_{A}^{ABC}, \lambda_{B}^{ABC}, \lambda_{C}^{ABC}, \lambda_{AB}^{ABC}, \lambda_{BC}^{ABC}, \lambda_{AC}^{ABC}, \lambda_{ABC}^{ABC}\right\}$$

and

$$\lambda_{\mathcal{P}_{\min}} = \left\{\lambda_{\varnothing}^{\varnothing}, \lambda_{A}^{A}, \lambda_{B}^{B}, \lambda_{C}^{C}, \lambda_{AB}^{AB}, \lambda_{BC}^{BC}, \lambda_{AC}^{AC}, \lambda_{ABC}^{ABC}\right\}, \tag{3.12}$$

respectively. (For notational simplicity we omitted the commas and braces in the sub- and superscripts.) Glonek (1996) considered a mixture of parameters taken from $\lambda_{\mathcal{P}_{\min}}$ and $\lambda_{\mathcal{P}_{\max}}$. However, the $\lambda_{\mathcal{P}}$s defined here are more general than these parameters.

3.3. *Smooth parameterizations* For an open set $\mathbf{B} \subseteq \mathbf{R}^k$, the parameter $\theta : \mathcal{F} \to \mathbf{B}$ is a t-dimensional ($1 \le t \le k$) *smooth parameterization* of $\mathcal{F}$ if it has the following properties:

R1: θ is a homeomorphism onto $\mathbf{B}$

R2: θ is twice continuously differentiable

R3: The Jacobian of θ has full rank t

The parameter θ is called smooth if R2 and R3 hold.

It will sometimes be convenient to work with sets of parameters rather than vectors. We will say that a distribution is parameterized by a set of parameters when it is parameterized by those parameters arranged in a vector. For a vector or set of parameters to be a smooth parameterization, redundant elements, i.e., parameters which are a function of others in the vector or set, must be removed. If θ is a vector or set of parameters, then $\tilde{\theta}$ denotes θ with the redundant elements removed.

A parameter $\lambda_{\mathcal{P}}$, with $\mathcal{P}$ complete, is not smooth because it contains redundant elements: summing $\lambda_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}_{\mathcal{L}})$ over an index (i.e., a coordinate of $\mathbf{i}_{\mathcal{L}}$) yields zero. To avoid this redundancy, define, for $i \leq p$, $I_i \in \mathcal{I}_i$ and $\tilde{\mathcal{I}}_i = \mathcal{I}_i \setminus \{I_i\}$, the subtable $\tilde{\mathcal{T}} = \times_{i=1}^{p} \tilde{\mathcal{I}}_i$ of $\mathcal{T}$. Then the parameter

$$\tilde{\lambda}_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}) : \mathcal{F} \to \mathbf{R}^{|\tilde{\mathcal{T}}|}$$

defined only for $\mathbf{i} \in \tilde{\mathcal{T}}$ does not contain redundant elements. We denote $\{\tilde{\lambda}_{\mathcal{L}}^{\mathcal{M}} : (\mathcal{L}, \mathcal{M}) \in \mathcal{P}\}$ by $\tilde{\lambda}_{\mathcal{P}}$.

For a hierarchical and complete $\mathcal{P}$, let $\mathbb{L}_i$ and $\mathbb{K}_i$ be defined as in (2.1), (2.2), and (2.3); for fixed $1 < i \leq s$ let $\mathbb{K}_i^c$ be the complement of $\mathbb{K}_i$ with respect to $\mathbb{P}(\mathcal{M}_i)$. Then $\mathbb{K}_i^c$ is a descending class of subsets of $\mathbb{P}(\mathcal{M}_i)$ in the sense that $H \subseteq \mathbb{K}_i^c$ implies that $H' \subseteq \mathbb{K}_i^c$ for every $H' \subseteq H$. On the other hand, $\mathbb{L}_i$ is an ascending class of subsets of $\mathbb{P}(\mathcal{M}_i)$ in the sense that $H \subseteq \mathbb{L}_i$ implies that $H' \subseteq \mathbb{L}_i$ for every $H' \supseteq H$. As $\mathcal{P}$ is complete and hierarchical, the classes $\mathbb{K}_i^c$ and $\mathbb{L}_i$ are complements to each other with respect to $\mathbb{P}(\mathcal{M}_i)$.

For $\mu \in \mathcal{F}$ and $\mathcal{P}_i = \{(\mathcal{L}, \mathcal{M}_i) : \mathcal{L} \in \mathbb{L}_i\}$, consider

$$\mu_{\mathcal{P}_i} = \begin{cases} \varnothing & i = 1 \\ \{\mu_K : K \in \mathbb{K}_i^c\} & i > 1 \end{cases}$$

and

$$\lambda_{\mathcal{P}_i} = \{\lambda_{\mathcal{L}}^{\mathcal{M}_i} : \mathcal{L} \in \mathbb{L}_i\}.$$

Then $\tilde{\lambda}_{\mathcal{P}_1}$ is the standard log-linear parameterization of the marginal distribution on $\mathcal{M}_1$ which, as is well-known, is smooth. If $i > 1$, then $\mu_{\mathcal{P}_i}$ contains the marginal distributions for a descending class of subsets of $\mathbb{P}(\mathcal{M}_i)$ and $\lambda_{\mathcal{P}_i}$ contains the (ordinary) log-linear parameters for the complement ascending class. As the $\mu_{\mathcal{P}_i}$ parameters are derived from a distribution μ and there can be no compatibility problems with the log-linear parameters in $\tilde{\lambda}_{\mathcal{P}_i}$, it is implied by general exponential family theory (see Barndorff-Nielsen, 1978, p. 112) that $\tilde{\mu}_{\mathcal{P}_i} \cup \tilde{\lambda}_{\mathcal{P}_i}$ is a mixed parameterization of the distribution on $\mathcal{M}_i$. This parameterization is smooth and its two components are variation independent. This is formulated in the next lemma.

LEMMA 2. *Suppose $\mathcal{P}$ is hierarchical and complete. If $1 \leq i \leq s$, then $\tilde{\mu}_{\mathcal{P}_i} \cup \tilde{\lambda}_{\mathcal{P}_i}$ is a smooth parameterization of the marginal distribution on $\mathcal{T}_{\mathcal{M}_i}$. If $i > 1$, then these two components are variation independent.*

As is well-known, if $\tilde{\mu}_{\mathcal{P}_i} \cup \tilde{\lambda}_{\mathcal{P}_i}$ has a given prescribed value, then $\mu_{\mathcal{M}_i}$ can be found by means of the iterative proportional fitting algorithm.

Lemma 2 can now be used to prove Theorem 2, which is the main result of this subsection.

THEOREM 2. *If $\mathcal{P}$ is hierarchical and complete, then $\tilde{\lambda}_{\mathcal{P}}$ is a smooth parameterization of $\mathcal{F}$.*

PROOF. If $s = 1$, then $\mathcal{P} = \mathcal{P}_1$ so the theorem follows directly from Lemma 2. If $s > 1$, the proof goes through a series of parameterizations of the distribution on the entire table $\mathcal{T}$, leading to the desired parameterization. By Lemma 2, $\mathcal{F}$ is smoothly parameterized by $\tilde{\mu}_{\mathcal{P}_s} \cup \tilde{\lambda}_{\mathcal{P}_s}$ and hence also by

$$(\{\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_{s-1}}\})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_s} \tag{3.13}$$

where $(.)^{\sim}$ stands for $\tilde{(.)}$.

Suppose $s \geq 3$. If, for $1 < i < s$, $\mathcal{F}$ is smoothly parameterized by

$$(\{\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_i}\})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_{i+1}} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s}$$

then, by Lemma 2, $\mathcal{F}$ is smoothly parameterized by

$$\begin{aligned}
&(\{\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_{i-1}}\} \cup \tilde{\mu}_{\mathcal{P}_i} \cup \tilde{\lambda}_{\mathcal{P}_i})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_{i+1}} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s} \\
&= (\{\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_{i-1}}\} \cup \tilde{\mu}_{\mathcal{P}_i})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_i} \cup \tilde{\lambda}_{\mathcal{P}_{i+1}} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s} \\
&= (\{\mu_{\mathcal{M}_1}, \dots, \mu_{\mathcal{M}_{i-1}}\})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_i} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s}.
\end{aligned}$$

Going through the above step for $i = s-1, \dots, 2$ gives that $\mathcal{F}$ is smoothly parameterized by

$$(\{\mu_{\mathcal{M}_1}\})^{\sim} \cup \tilde{\lambda}_{\mathcal{P}_2} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s} = \{\mu_{\mathcal{M}_1}\} \cup \tilde{\lambda}_{\mathcal{P}_2} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s}. \tag{3.14}$$

If $s = 2$, then (3.13) is identical to (3.14). Thus, $\mathcal{F}$ is smoothly parameterized by (3.14) for all $s > 1$. By Lemma 2, the marginal distribution on $\mathcal{M}_1$ is smoothly parameterized by $\tilde{\lambda}_{\mathcal{P}_1}$, so $\mathcal{F}$ is smoothly parameterized by

$$\tilde{\lambda}_{\mathcal{P}} = \tilde{\lambda}_{\mathcal{P}_1} \cup \tilde{\lambda}_{\mathcal{P}_2} \cup \dots \cup \tilde{\lambda}_{\mathcal{P}_s}$$

and this completes the proof.

A proof based on the same idea, namely the sequence of mixed parameterizations, was used by Kauermann (1997) to show that $\tilde{\lambda}_{\mathcal{P}_{\min}}$ is invertable. However, his proof uses a specific recursive property of $\lambda_{\mathcal{P}_{\min}}$ and appears to be difficult to apply to the general case described here.

Theorem 2 together with Lemma 1 implies that every parameter $\tilde{\lambda}_{\mathcal{P}}$, based on a hierarchical $\mathcal{P}$, can be completed to yield a smooth parameterization, i.e.,

Corollary 1. *If $\mathcal{P}$ is hierarchical, then $\tilde{\lambda}_{\overline{\mathcal{P}}}$ is a smooth parameterization of the distributions in $\mathcal{F}$.*

Theorem 3 demonstrates that for certain non-hierarchical $\mathcal{P}$s, the parameter $\tilde{\lambda}_{\mathcal{P}}$ is not smooth.

Theorem 3. *Suppose $\{(\mathcal{L}, \mathcal{M}), (\mathcal{L}, \mathcal{N})\} \subseteq \mathcal{P}$ for certain $\mathcal{L} \subseteq \mathcal{M}, \mathcal{N} \subseteq \mathcal{V}$. Then $\tilde{\lambda}_{\mathcal{P}}$ is not smooth.*

Proof. For $\mathcal{L} \subseteq \mathcal{V}$ and arbitrary $\mathbf{i}, \mathbf{j} \in \mathcal{T}$, define

$$\begin{aligned}
d_{\varnothing}(\mathbf{i}_{\varnothing}, \mathbf{j}_{\varnothing}) &= 1 \\
d_{\mathcal{L}}(\mathbf{i}_{\mathcal{L}}, \mathbf{j}_{\mathcal{L}}) &= I(\mathbf{i}_{\mathcal{L}} = \mathbf{j}_{\mathcal{L}})\, |\mathcal{T}_{\mathcal{L}}| - \sum_{\mathcal{L}' \subset \mathcal{L}} d_{\mathcal{L}'}(\mathbf{i}_{\mathcal{L}'}, \mathbf{j}_{\mathcal{L}'})
\end{aligned}$$

where $I(.)$ is the indicator function, giving one if the argument is true, zero otherwise. Then it can be verified that for arbitrary $\mathcal{U}$ for which $\mathcal{L} \subseteq \mathcal{U} \subseteq \mathcal{V}$,

$$\frac{\partial \lambda_{\mathcal{L}}^{\mathcal{U}}(\mathbf{i}_{\mathcal{L}})}{\partial \mu(\mathbf{j})} = \frac{d_{\mathcal{L}}(\mathbf{i}_{\mathcal{L}}, \mathbf{j}_{\mathcal{L}})}{|\mathcal{T}_{\mathcal{U}}|\, \mu_{\mathcal{U}}(\mathbf{j}_{\mathcal{U}})}.$$

It follows that

$$\left.\frac{\partial \lambda_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}_{\mathcal{L}})}{\partial \mu(\mathbf{j})}\right|_{\mu(\mathbf{j})=|\mathcal{T}|^{-1}} = \left.\frac{\partial \lambda_{\mathcal{L}}^{\mathcal{N}}(\mathbf{i}_{\mathcal{L}})}{\partial \mu(\mathbf{j})}\right|_{\mu(\mathbf{j})=|\mathcal{T}|^{-1}} = d_{\mathcal{L}}(\mathbf{i}_{\mathcal{L}}, \mathbf{j}_{\mathcal{L}}).$$

Since $\mathbf{j}$ was taken arbitrarily, the Jacobians of $\lambda_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}_{\mathcal{L}})$ and $\lambda_{\mathcal{L}}^{\mathcal{N}}(\mathbf{i}_{\mathcal{L}})$ are identical when evaluated at the uniform distribution $\mu(\mathbf{k}) = |\mathcal{T}|^{-1}$ ($\mathbf{k} \in \mathcal{T}$). Their contributions to the Jacobian of $\tilde{\lambda}_{\mathcal{P}}$ are therefore also identical. Hence, $\tilde{\lambda}_{\mathcal{P}}$ is not of full rank for all distributions in $\mathcal{F}$, and is therefore not smooth because R3 is violated.

3.4. *Variation independence* Suppose $\theta = (\theta_1, \ldots, \theta_t)$ is a t-dimensional parameter. Then θ is variation independent if $\mathcal{R}(\theta) = \mathcal{R}(\theta_1) \times \ldots \times \mathcal{R}(\theta_t)$, where $\mathcal{R}(.)$ denotes the range of a function. It is well-known that the ordinary log-linear parameters are variation independent.

THEOREM 4. *Let $\mathcal{P}$ be hierarchical and complete. Then $\tilde{\lambda}_{\mathcal{P}}$ is variation independent if and only if $\mathcal{P}$ is ordered decomposable.*

PROOF. "$\Leftarrow$": The proof goes through the steps in the proof of Theorem 2 in reverse order. Let $\mathcal{M}_1, \ldots, \mathcal{M}_s$ be a hierarchical and ordered decomposable ordering of the elements of $\mathbb{M}$. Let $\mathbf{a} \in \mathbf{R}^{|\tilde{\lambda}_{\mathcal{P}}|}$ be an arbitrary real vector. It will be shown that a μ exists, for which $\tilde{\lambda}_{\mathcal{P}} = \mathbf{a}$.

First note that, since $\tilde{\lambda}_{\mathcal{P}_1}$ is a smooth parameterization of the distribution on the $\mathcal{M}_1$ marginal, $\mu_{\mathcal{M}_1}$, and with it also $\mu_{\mathcal{P}_1}$ can be constructed. If $\mu_{\mathcal{P}_i}$ is available for some $i = 1, \ldots, s$, then, since $\lambda_{\mathcal{P}_i}$ is prescribed, $\mu_{\mathcal{M}_i}$ can be constructed by Lemma 2. From $\mu_{\mathcal{M}_1}, \ldots, \mu_{\mathcal{M}_i}$ ($i < s$), $\mu_{\mathcal{P}_{i+1}}$ can also be deducted implying that $\mu_{\mathcal{M}_s} = \mu$ can also be constructed.

"$\Rightarrow$": Suppose $\mathcal{P}$ is complete but not ordered decomposable. Then for a hierarchical ordering $\mathcal{M}_1, \ldots, \mathcal{M}_s$, there is an $i \leq s$ such that the set of maximal elements of $\{\mathcal{M}_1, \ldots, \mathcal{M}_j\}$ is decomposable for all $j < i$ but the set of maximal elements of $\{\mathcal{M}_1, \ldots, \mathcal{M}_i\}$ is not decomposable. But then, by Theorem 1, there is a $(\mu_{\mathcal{M}_1}, \ldots, \mu_{\mathcal{M}_i})$ which is weakly compatible but not strongly compatible. By completeness of $\mathcal{P}$, there exists a marginal log-linear parameter $\tilde{\lambda}_{\mathcal{P}}$ which yields these marginals. Therefore, this $\tilde{\lambda}_{\mathcal{P}}$ is not variation independent.

Theorem 4 is a solution to a problem encountered by several authors. Liang et al. (1992), Glonek and McCullagh (1995), and Kauermann (1997) noted that prescribing values to certain marginal log-linear parameters restricts the range of others. The theorem shows that, provided $\mathcal{P}$ is ordered decomposable, $\tilde{\lambda}_{\mathcal{P}}$ may be freely prescribed. This eases the interpretation of $\tilde{\lambda}_{\mathcal{P}}$ and may have some computational advantages, but it does not mean a researcher should refrain from using a $\tilde{\lambda}_{\mathcal{P}}$ based on a $\mathcal{P}$ which is not ordered decomposable if this is required by the substantive problem at hand.

Note that if $\mathcal{P}$ is neither complete nor ordered decomposable, conditions for variation independence of $\tilde{\lambda}_{\mathcal{P}}$ are yet unknown.

It follows from the proof of Theorem 4, that given a strongly compatible prescribed value of $\lambda_{\mathcal{P}}$, repeated application of the iterative proportional fitting procedure yields the corresponding distribution μ. McCullagh and Nelder (1989, Exercise 6.8) and Kauermann (1997) demonstrated that this is so for $\lambda_{\mathcal{P}_{\min}}$.

The following example demonstrates that $\tilde{\lambda}_{\mathcal{P}}$ cannot always be assigned arbitrary values if $\mathcal{P}$ is hierarchical but not ordered decomposable.

EXAMPLE 4. *Consider the parameters of the multivariate logistic transform (3.12), and let*

$$\lambda_{\varnothing}^{\varnothing} = \log 8 \tag{3.15}$$

$$\lambda_A^A(1) = \lambda_B^B(1) = \lambda_C^C(1) = 0 \tag{3.16}$$

$$\lambda_{AB}^{AB}(1,1) = \lambda_{BC}^{BC}(1,1) = \log(3/8), \qquad \lambda_{AC}^{AC}(1,1) = \log(1/8). \tag{3.17}$$

Equation (3.15) constrains the sample size to be equal to 8, (3.16) ensures the one-dimensional marginal frequencies are equal in each category of every variable, and (3.17) sets the two-dimensional marginal odds ratios to 9, 9 and 1/9, respectively. This yields the 2-dimensional marginal tables AB, BC, and AC

$$A\overset{B}{\begin{array}{|c|c|}\hline 3&1\\\hline 1&3\\\hline\end{array}}\quad B\overset{C}{\begin{array}{|c|c|}\hline 3&1\\\hline 1&3\\\hline\end{array}}\quad A\overset{C}{\begin{array}{|c|c|}\hline 1&3\\\hline 3&1\\\hline\end{array}},$$

respectively. It is straightforward to verify that, although the parameters in (3.15) to (3.17) are weakly compatible, there is no distribution in $\mathcal{F}$ with these marginal parameters. For example, it can be checked that the product-moment correlation matrix for the three variables is not positive definite.

Note that replacing, for example, λ_{AC}^{AC} in the multivariate logistic transform by λ_{AC}^{ABC}, yields the parameter with variation independent components

$$(\lambda_{\varnothing}^{\varnothing}, \lambda_A^A, \lambda_B^B, \lambda_C^C, \lambda_{AB}^{AB}, \lambda_{BC}^{BC}, \lambda_{AC}^{ABC}, \lambda_{ABC}^{ABC})$$

which, for any weakly compatible assignment of real values, yields a distribution in $\mathcal{F}$.

4. Log-affine and log-linear marginal models A subset $\mathcal{G} \subseteq \mathcal{F}$ is called a *model* in $\mathcal{F}$. For given $\mathcal{P}$, let $\mathcal{H}$ be a nonempty linear subspace of $\mathbf{R}^{\dim(\tilde{\lambda}_{\mathcal{P}})}$ and let $\mathbf{q} \in \mathbf{R}^{\dim(\tilde{\lambda}_{\mathcal{P}})}$. A *log-affine marginal model* $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H}) \subseteq \mathcal{F}$ is defined by the restriction

$$\mu \in M_{\mathcal{P}}(\mathbf{q}, \mathcal{H}) \Leftrightarrow \tilde{\lambda}_{\mathcal{P}} \in \mathbf{q} + \mathcal{H}.$$

Note that $\mathbf{q} \in \mathbf{q} + \mathcal{H}$. Log-affine marginal models generalize the ordinary log-linear models, log-affine models (Haberman, 1974; Rudas & Leimer, 1992; Lauritzen, 1996), the multivariate logistic models of McCullagh and Nelder (1989) and Glonek and McCullagh (1995), and the "mixture" of these models considered by Glonek (1996). The log-affine marginal model with $\mathbf{q} = \mathbf{0}$, i.e., $M_{\mathcal{P}}(\mathbf{0}, \mathcal{H})$ is called a *log-linear marginal model.* Applications of log-linear marginal models have been considered by McCullagh and Nelder (1989), Liang et al. (1992), Agresti and Lang (1993), Becker (1994), Lang and Agresti (1994), Croon, Bergsma, and Hagenaars (2000), and various others. Example 5 shows some of the variety of log-affine marginal models and demonstrates that there are interesting models which are log-affine but not multivariate logistic.

EXAMPLE 5. *Let $\mathcal{V} = \{A, B, C, D, V, W\}$. The log-linear marginal model asserting that the marginal association between A and B equals the one between C and D can be formulated as*

$$\lambda_{AB}^{AB}(i, j) = \lambda_{CD}^{CD}(i, j).$$

This model is also multivariate logistic. On the other hand, the model specifying that the marginal association between A and B given *V equals the one between C and D* given *W can be formulated as*

$$\lambda_{ABV}^{ABV}(i, j, k) = \lambda_{CDW}^{CDW}(i, j, k)$$
$$\lambda_{AB}^{ABV}(i, j) = \lambda_{CD}^{CDW}(i, j).$$

This is, in fact, a log-linear marginal model, but not an ordinary log-linear or a multivariate logistic one, nor a mixture of the latter two.

The inclusion of covariates can be done as follows. Suppose $\mathcal{V} = \cup_{t=1}^{s} \mathcal{M}_t$, $|\mathcal{M}_t| = 2$. If $\mu_{\mathcal{M}_t}$ is the bivariate marginal distribution at time point t, a trend in the association may be specified by the log-linear marginal model

$$\lambda_{\mathcal{M}_t}^{\mathcal{M}_t}(i, j) = x_t \beta_{ij},$$

where the β_{ij} are unknowns and x_t is constant for every time point t.

As can be seen from the examples, log-affine marginal models can be constructed in the log-linear tradition. One can select those marginal distributions which are of interest, and then use the marginal log-linear parameters to build models, either for modeling relations *between* various marginal distributions, or for modeling the relations between certain variables *within* a marginal distribution.

Log-affine marginal models can also be used to describe the sampling procedure that is used. That is, for certain $\mathcal{P}$, $\mathbf{q}$, and $\mathcal{H}$, the sampling procedure may be designed such that the observed distribution is an element of $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$. For example, the sampling design may be such that various marginal distributions are fixed. This provides a unified treatment of the substantive model and the sampling procedure, which is not possible with ordinary log-linear models. Example 6 shows that the sampling procedures with the total sample size fixed (multinomial sampling), one marginal fixed (stratified or product multinomial sampling), and two marginals fixed can be described by log-affine marginal models.

Example 6. *Suppose $A, B \in \mathcal{V}$ have index sets $\mathcal{I}_A$ and $\mathcal{I}_B$, respectively. Let N_A and N_B be compatible marginal frequency distributions for variables A and B, respectively, i.e., there is a frequency $N > 0$ such that $\sum N_A(i) = \sum N_B(j) = N$.*

For the multinomial sampling model*, the total number of counts N is fixed by design and is specified by the restriction*

$$\lambda_\varnothing^\varnothing = \log N.$$

This can be seen by noting that $\lambda_\varnothing^\varnothing = \log \mu_\varnothing$. To fix the marginal distribution of variable A to N_A, the additional restriction

$$\lambda_\varnothing^A + \lambda_A^A(i) = \log N_A(i)$$

is needed, since $\lambda_\varnothing^A + \lambda_A^A(i) = \log \mu_A(i)$ (see (3.10) and (3.11)). This is called the stratified sampling model *with stratifying variable A. To fix, additionally, the distribution of B to N_B, the further restriction*

$$\lambda_\varnothing^B + \lambda_B^B(i) = \log N_B(j)$$

is needed.

If $\mathcal{F}$ is an exponential family, a model $\mathcal{G} \subseteq \mathcal{F}$ is *curved exponential* if it has a smooth parameterization R1–R3. Its dimension is the local dimension of the parameter space, t in R1–R3. (See also Lauritzen, 1996, p. 272.)

Theorem 5. *If $\mathcal{P}$ is hierarchical and complete then a nonempty log-affine marginal model $M_\mathcal{P}(\mathbf{q}, \mathcal{H})$ is a curved exponential model of dimension* $\dim(\mathcal{H})$*.*

Proof. By Theorem 2, $\tilde{\lambda}_\mathcal{P}$ is a smooth parameterization of $\mathcal{F}$. Then an affine combination of its coordinates is a smooth parameterization of $M_\mathcal{P}(\mathbf{q}, \mathcal{H})$ with dimension $\dim(\mathcal{H})$.

The restriction to complete $\mathcal{P}$ in Theorem 5 is without any loss of generality. In particular, if $\mathcal{P}$ is hierarchical but incomplete, then by Lemma 1 $\overline{\mathcal{P}} \supset \mathcal{P}$ is hierarchical and complete. Then $\tilde{\lambda}_\mathcal{P} \in \mathbf{q} + \mathcal{H}$ is equivalent to

$$\tilde{\lambda}_{\overline{\mathcal{P}}} = \begin{pmatrix} \tilde{\lambda}_\mathcal{P} \\ \tilde{\lambda}_{\overline{\mathcal{P}} \setminus \mathcal{P}} \end{pmatrix} \in \begin{pmatrix} \mathbf{q} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathcal{H} \\ \mathbf{0}^\perp \end{pmatrix}$$

where $\mathbf{0}$ is a vector of zeroes with the same dimension as $\tilde{\lambda}_{\overline{\mathcal{P}} \setminus \mathcal{P}}$, and $\mathbf{0}^\perp$ is its null space.

Example 7 shows that Theorem 5 does not necessarily hold if the hierarchy condition is violated.

Example 7. *Let $\mathcal{V} = \{A, B, C\}$ be a set of dichotomous variables, and let*

$$\mathcal{P} = \{(\{A, B\}, \{A, B\}), (\{A, B\}, \{A, B, C\}), (\{A, B, C\}, \{A, B, C\})\}.$$

Then $\mathcal{P}$ is not hierarchical (see also Example 1). We show that certain linear restrictions on $\lambda_\mathcal{P}$ yield a model which is not curved exponential.

Marginal independence of A and B is specified as

$$\lambda_{AB}^{AB}(1,1) = 0. \tag{4.18}$$

Conditional independence of A and B given C is specified as

$$\lambda_{ABC}^{ABC}(1,1,1) = \lambda_{AB}^{ABC}(1,1) = 0. \tag{4.19}$$

Let $M_{\mathcal{P}}(\mathbf{0},\mathcal{H})$ be the log-linear marginal model defined by the linear restrictions (4.18) and (4.19). Dawid (1980) showed that $M_{\mathcal{P}}(\mathbf{0},\mathcal{H})$ is equivalent to A being independent of both B and C, or B being independent of both A and C (or both). In terms of prescriptions for log-linear parameters, this is

$$\lambda_{ABC}^{ABC}(1,1,1) = \lambda_{AB}^{ABC}(1,1) = \lambda_{AC}^{ABC}(1,1) = 0$$

or

$$\lambda_{ABC}^{ABC}(1,1,1) = \lambda_{AB}^{ABC}(1,1) = \lambda_{BC}^{ABC}(1,1) = 0.$$

Since $M_{\mathcal{P}}(\mathbf{0},\mathcal{H})$ is the union of two distinct, intersecting models, it cannot be the smooth image of an open set. Hence, $M_{\mathcal{P}}(\mathbf{0},\mathcal{H})$ is not curved exponential.

The problem of the determination of the dimension of a marginal model for categorical data was posed by Lang and Agresti (1994). They solved the problem for the case when restrictions are placed on certain ordinary log-linear parameters and on certain marginals. Theorem 5 gives a general solution to the dimension problem for log-affine marginal models based on hierarchical $\mathcal{P}$, provided nonemptiness can be established.

For $a, b \in \mathbf{R}^k$, a continuous function $f : [0,1] \to \mathbf{R}^k$ such that $f(0) = a$ and $f(1) = b$ is a *path* between a and b. A subset of $\mathbf{R}^k$ is *connected* if there is a path between any two points in the subset. We have:

THEOREM 6. *If $\mathcal{P}$ is ordered decomposable, the log-affine marginal model $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$ is nonempty and connected.*

PROOF. Assume that $\mathcal{P}$ is hierarchical and complete. (See the comment after Theorem 5.) Then by Theorem 2, $\tilde{\lambda}_{\mathcal{P}}$ is a smooth parameterization of $\mathcal{F}$. Hence, $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$ is homeomorph with $\mathbf{q}+\mathcal{H}$, which is nonempty and connected. Thus, $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$ is also nonempty and connected. The argument after Theorem 5 shows that the proof extends to $\mathcal{P}$s which are not complete.

In general, it may be difficult to check whether or not a log-affine model $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$ is empty. However, all examples of marginal models which have been described in the literature referred to in this article are, in fact, of the form $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$, with $\mathcal{P}$ ordered decomposable. Example 4 demonstrates that there may be no distribution in $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$ if $\mathcal{P}$ is not ordered decomposable. An open question is exactly which of the latter types of models exist. Another open problem which remains is whether or not there are log-linear or log-affine marginal models $M_{\mathcal{P}}(\mathbf{q},\mathcal{H})$, with $\mathcal{P}$ not ordered decomposable, which are not connected.

Theorem 7 shows that nonemptiness of marginal models which are log-linear is generally guaranteed.

THEOREM 7. *If there is at most one $\mathcal{M} \subseteq \mathbb{P}(\mathcal{V})$ such that $(\varnothing, \mathcal{M}) \in \mathcal{P}$, then the log-linear marginal model $M_{\mathcal{P}}(\mathbf{0}, \mathcal{H})$ is nonempty.*

PROOF. Suppose there is no $\mathcal{N} \subseteq \mathbb{P}(\mathcal{V})$ such that $(\varnothing, \mathcal{N}) \in \mathcal{P}$. Then, for $\mathbf{i} \in \mathcal{T}$, any uniform distribution $\mu(\mathbf{i}) = c$ $(c > 0)$ yields $\lambda_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}_{\mathcal{L}}) = 0$ $((\mathcal{L}, \mathcal{M}) \in \mathcal{P})$, so the uniform distribution is in the log-linear marginal model $M_{\mathcal{P}}(\mathbf{0}, \mathcal{H})$.

If $(\varnothing, \mathcal{N}) \in \mathcal{P}$, then, for $\mathbf{i} \in \mathcal{T}$, the uniform distribution $\mu(\mathbf{i}) = 1/|\mathcal{T}_{\mathcal{V}\setminus\mathcal{N}}|$ yields $\mu(\mathbf{i}_{\mathcal{N}}) = 1$. Hence, $\lambda_{\varnothing}^{\mathcal{N}}(\mathbf{i}_{\varnothing}) = 0$, and, by the above argument, all other marginal log-linear parameters are zero also, and so the model is nonempty.

Collapsibility conditions for contingency tables can be specified using log-linear marginal models, in particular as a model $M_{\mathcal{P}}(\mathbf{0}, \mathcal{H})$, with $\mathcal{P}$ non-hierarchical. Whittemore (1978) defined a table $\mathcal{T}$ to be *collapsible* onto the marginal table $\mathcal{T}_{\mathcal{M}}$ with respect to $\mathcal{L} \subseteq \mathcal{M}$ if

$$\lambda_{\mathcal{L}}^{\mathcal{V}}(\mathbf{i}_{\mathcal{L}}) = \lambda_{\mathcal{L}}^{\mathcal{M}}(\mathbf{i}_{\mathcal{L}}) \tag{4.20}$$

for all $\mathbf{i} \in \mathcal{T}$. This means that the complete table $\mathcal{T}$ and the marginal table $\mathcal{T}_{\mathcal{M}}$ contain the same amount of information about the interactions between the variables in $\mathcal{L}$. If, additionally,

$$\lambda_{\mathcal{K}}^{\mathcal{V}}(\mathbf{i}_{\mathcal{K}}) = 0 \tag{4.21}$$

for all $\mathcal{L} \subset \mathcal{K} \not\subseteq \mathcal{M}$, then $\mathcal{T}$ is said to be *strictly collapsible* onto $\mathcal{T}_{\mathcal{M}}$ with respect to $\mathcal{L}$. This means that the association between the variables in $\mathcal{L}$ is the same in marginal table $\mathcal{T}_{\mathcal{M}}$ as in table $\mathcal{T}$ conditionally on any (subset of) variable(s) not in $\mathcal{M}$. Theorem 7 shows that these collapsibility conditions can always be imposed upon an ordinary log-linear model or, more generally, a log-linear marginal model. By Theorem 3, the parameters restricted in (4.20) are not smooth. As a result, collapsibility conditions generally do not define a curved exponential family. Example 7 provides an example of strict collapsibility of $\mathcal{T}_{ABC}$ onto $\mathcal{T}_{AB}$ with respect to variables A and B, combined with marginal independence between A and B (and therefore also conditional independence on C).

More generally than the log-linear or log-affine marginal models discussed in this section, non-linear models for marginal log-linear parameters can be considered. In fact, the theorems of this section do not depend on the linearity of $\mathcal{H}$; if $\mathcal{H}$ is homeomorph with a linear subspace and contains the origin, all the theorems still hold. Interesting nonlinear models for categorical data include the row and column effects model (Goodman, 1979). This particular model has been used in the marginal modeling framework by Colombi (1998).

5. Asymptotic maximum likelihood theory Let x be a random count variable defined on $\mathcal{T}$, i.e., $x \in \mathbf{N}^{|\mathcal{T}|}$, where $\mathbf{N} = \{0, 1, \ldots\}$. It is assumed that $x(\mathbf{i})$ $(\mathbf{i} \in \mathcal{T})$ has a Poisson distribution with mean $\mu(\mathbf{i})$, where μ is the expectation of x. The MLE of μ under a log-affine marginal model $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$ is the value which maximizes the likelihood of x over $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$, and is denoted as $\hat{\mu}_x$.

Let $x_1, \ldots, x_n$ be independent and distributed identically as x and let $\bar{x}_n = \sum_{i=1}^n x_i/n$ be their average. That is, for each cell $\mathbf{i}$ in table $\mathcal{T}$, $\bar{x}_n$ contains the average count of the corresponding counts of the x_i. Theorem 8 describes the behavior of $\hat{\mu}_{\bar{x}_n}$ as $n \to \infty$, and follows directly from Theorem 5 and theory of exponential family distributions (Barndorff-Nielsen, 1978; Lauritzen, 1996).

THEOREM 8. *For a hierarchical $\mathcal{P}$ suppose $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$ is a nonempty log-affine marginal model containing μ. Then the probability that $\hat{\mu}_{\bar{x}_n}$ exists uniquely and is stationary in $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$ tends to 1 as n increases to infinity, and $\hat{\mu}_{\bar{x}_n}$ has an asymptotic multivariate normal distribution with mean μ.*

Lauritzen (1996, Section D.2.1) shows how to calculate the asymptotic covariance matrix of $\hat{\mu}_{\bar{x}_n}$. It is omitted here for ease of exposition, but can be found in Lang (1996; see also Bergsma, 1997, Appendix A.1). It follows from the asymptotic normality of $\hat{\mu}_{\bar{x}_n}$ that standard goodness-of-fit statistics, such as the likelihood ratio or Pearson's chi-square, have an asymptotic chi-square distribution. If $\mathcal{P}$ is complete, the latter has $|\mathcal{T}| - \dim(\mathcal{H})$ degrees of freedom, otherwise this number must be reduced by the dimension of $\tilde{\lambda}_{\overline{\mathcal{P}} \setminus \mathcal{P}}$ (see the comment after Theorem 5).

Theorem 8 can be generalized to certain cases where the sampling design is more restrictive than in the simple Poisson sampling design considered above. In particular, the sampling design may be such that $x \in M_{\mathcal{P}'}(\mathbf{q}', \mathcal{H}')$ for some nonempty log-affine marginal model $M_{\mathcal{P}'}(\mathbf{q}', \mathcal{H}')$. For example, for multinomial sampling with fixed $N > 0$, $M_{\mathcal{P}'}(\mathbf{q}', \mathcal{H}') = \{w \in \mathbf{N}^{|\mathcal{T}|} : \sum_{\mathbf{i} \in \mathcal{T}} w(\mathbf{i}) = N\}$. Note that one multinomial sample of size N is considered as a single observation of x. Generally, $M_{\mathcal{P}'}(\mathbf{q}', \mathcal{H}')$ may be used to describe sampling designs with certain marginals fixed. (See Example 6. However, note that other designs are also possible.) For Theorem 8 to generalize, the restriction must be made that $M_{\mathcal{P}'}(\mathbf{q}', \mathcal{H}') \cap M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$ contains μ and is a curved exponential family. By Theorem 5, it is sufficient that the intersection of the statistical and sampling models is nonempty and that $\mathcal{P} \cup \mathcal{P}'$ is hierarchical.

It should be noted that without the hierarchy condition Theorem 8 is not true. For example, the model discussed in Example 7 is based on a non-hierarchical $\mathcal{P}$. If the special case of the model in which A, B, and C are mutually independent is true, then with certain probability p_n, $\hat{\mu}_{\bar{x}_n}$ is in model $A \perp\!\!\!\perp C|B$ and with certain probability q_n, $\hat{\mu}_{\bar{x}_n}$ is in model $B \perp\!\!\!\perp C|A$. It can be shown that $\lim_{n\to\infty} p_n = p^*$ and $\lim_{n\to\infty} q_n = 1 - p^*$ for some fixed $0 < p^* < 1$. Hence, the asymptotic distribution of $\hat{\mu}_{\bar{x}_n}$ is not multivariate normal. Of particular importance is that the conditional likelihood ratio statistic for testing mutual independence of A, B, and C against the alternative that A and B are both marginally independent and conditionally independent given C is asymptotically distributed as the minimum of two chi-square statistics, with possibly different degrees of freedom.

As stated above, Theorem 8 follows from Theorem 5 and exponential family theory. For classes of distributions more general than the exponential family, Aitchison and Silvey (1958) gave conditions that the properties of Theorem 8 hold for maximum likelihood estimates of parameters subject to constraints. With $\mathbf{C}$ and $\mathbf{A}$ matrices satisfying certain regularity conditions, and $\boldsymbol{\mu}$ the vector of expected cell frequencies, Lang (1996) considered the class of models which can be described by the equation $\mathbf{C}\log\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, which includes the class of log-affine marginal models. He verified Aitchison and Silvey's conditions for these models under Poisson and multinomial sampling, and under the assumption that the Jacobian of $\mathbf{C}\log\mathbf{A}\boldsymbol{\mu}$ has full rank. By Theorem 2, the latter assumption is true for models $M_{\mathcal{P}}(\mathbf{q}, \mathcal{H})$, with $\mathcal{P}$ hierarchical, i.e., Aitchison and Silvey's conditions, instead of exponential family theory, can also be used to prove Theorem 8.

As a final note, stationarity of the maximum likelihood estimate $\hat{\mu}_{\bar{x}_n}$ in a log-affine marginal model is important because it allows a gradient algorithm to be applied to find it. Various authors have described different gradient algorithms, for example, Haber (1985), Lang and Agresti (1994), Molenberghs and Lesaffre (1994), Glonek and McCullagh (1995), Bergsma (1997), and Colombi and Forcina (2000).

Acknowledgments. The authors wish to thank the reviewers and the Associate Editor for their thorough reading of the first version of the manuscript and for several helpful comments.

REFERENCES

Agresti, A., & Lang, J. B. (1993). A proportional odds model with subject-specific effects for repeated ordered responses. *Biometrika*, *80*, 527-534.

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

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

Becker, M. P. (1994). Analysis of repeated categorical measurements using models for marginal distributions: an application to trends in attitudes on legalized abortion. In P. V. Marsden (Ed.), *Sociological methodology.* Oxford, UK: Blackwell.

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

Colombi, R. (1998). A multivariate logit model with marginal canonical association. *Comm. Stat., Theory & Methods*, *27*, 2953-2972.

Colombi, R., & Forcina, A. (2000). Marginal regression models for the analysis of positive association of ordinal response variables. *Submitted.*

Croon, M. A., Bergsma, W. P., & Hagenaars, J. A. (2000). Analyzing change in categorical variables by generalized log-linear models. *Sociol. Methods & Res.*, *Vol. 29 No 2*, 195–229.

Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Stat.*, *8*, 552-539.

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

Glonek, G. J. N. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, *83*(1), 15-28.

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

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statis. Assoc.*, *74*, 537-552.

Haber, M. (1985). Maximum likelihood methods for linear and loglinear models in categorical data. *Comput. Statist. Data anal.*, *3*, 1-10.

Haberman, S. J. (1974). *The analysis of frequency data.* Chicago: University of Chicago Press.

Kauermann, G. (1997). A note on multivariate logistic models for contingency tables. *Austr. J. Stat.*, *39*, 261-276.

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

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

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

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. Roy. Statist. Soc. Ser. B*, *54*, 3-40.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman and Hall.

Molenberghs, G., & Lesaffre, E. (1994). Marginal modelling of multivariate categorical data. *J. Am. Statist. Assoc.*, *89*, 633-644.

Rudas, T., & Leimer, H. G. (1992). Analysis of contingency tables with known conditional odds ratios or known log-linear parameters. In B. Francis, G. U. H. Seeberg, P. G. M. van der Heijden, & W. Jansen (Eds.), *Statistical modelling* (pp. 312–322). Elsevier.

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. B*, *40*, 328-340.

DEPARTMENT OF METHODOLOGY
FAC. OF SOCIAL AND BEHAVIORAL SCIENCES
TILBURG UNIVERSITY
PO BOX 90153
5000 LE TILBURG, THE NETHERLANDS

DEPARTMENT OF STATISTICS
INSTITUTE OF SOCIOLOGY
EÖTVÖS LORÁND UNIVERSITY
PÁZMÁNY PÉTER SÉTÁNY 1/A
H-1117 BUDAPEST, HUNGARY