Computing Maximum Likelihood Estimates in Log-Linear Models

Alessandro Rinaldo*
Department of Statistics
Carnegie Mellon University

June, 2006

Abstract

We develop computational strategies for extended maximum likelihood estimation, as defined in Rinaldo (2006), for general classes of log-linear models of widespred use, under Poisson and product-multinomial sampling schemes. We derive numerically efficient procedures for generating and manipulating design matrices and we propose various algorithms for computing the extended maximum likelihood estimates of the expectations of the cell counts. These algorithms allow to identify the set of estimable cell means for any given observable table and can be used for modifying traditional goodness-of-fit tests to accommodate for a nonexistent MLE. We describe and take advantage of the connections between extended maximum likelihood estimation in hierarchical log-linear models and graphical models.

Contents

*Email: arinaldo@stat.cmu.edu

1 Introduction

Log-linear models are a powerful statistical tool for the analysis of categorical data and their use has increased greatly over the past two decades with the compilation and distribution of large,

and very often sparse, data bases, in the social and medical sciences as well as in machine learning applications. The growing interest in analyzing massive, sparse databases has dramatically exposed a critical and, until very recently, unresolved issue with log-linear modeling: the possibility that the presence of sampling zeros compromises the feasibility and correctness of statistical inference. Sampling zeros, i.e. cells containing zero counts, arise as a consequence of the random nature of the sampling mechanism itself and occur frequently, but not exclusively, in tables with a relatively large number of cells compared to the sample size. Sampling zeros may be thought of as missing bits of information. When they occur in specific patterns inside the table, the maximum of the likelihood function occurs at the boundary of the parameter space where some subset of the expected values are also zero. In such cases, we say that the Maximum Likelihood Estimate (MLE) of the cell mean vector does not exist.

The MLE of the expected value of the vector of observed counts plays a fundamental role for assessment of fit, model selection and interpretation. The existence of the MLE is essential for the usual derivation of large sample χ^2 approximations to numerous measures of goodness of fit (Bishop et al., 1975; Cressie and Read, 1988; Agresti, 2002) which are used in testing and model selection. If the distribution of the goodness-of-fit statistic is instead derived from the "exact distribution," i.e., the conditional distribution given the sufficient statistics, namely the margins, it is still necessary in most cases to have an MLE or some similar type of estimator in order to quantify the discrepancy of the the observed data from the fitted values. In addition, the existence of the MLE is essential to the derivation of the limiting distribution in the double-asymptotic approximations for the likelihood ratio and Pearson's χ^2 statistics for tables in which both the sample size and the number of cells are allowed to grow unbounded (Cressie and Read, 1988). If the MLE is not defined, the inferential procedures mentioned above may not be applicable or, at a minimum, require adjusting the degrees of freedom.

The problem of nonexistence of the MLE has long been known to be related to the presence of zero cell counts in the observed table (see, in particular, Haberman, 1974; Bishop et al., 1975). Even if a zero entry in the margins is a sufficient condition for the nonexistence of the MLE, little has been known about other "pathological" cases of tables with positive margins but where the MLE still does not exist. The most famous, and until recently, the only published example of this kind is the 2^3 table and the model of no-second-order interaction described by Haberman (1974) (see Example 6.1 below). Although Haberman (1974) gave necessary and sufficient conditions for the existence of the MLE, his characterization is nonconstructive in the sense that it does not directly lead to implementable numerical procedures and also fails to suggest alternative methods of inference for the case of an undefined MLE. Despite these deficiencies, Haberman's results have not been improved or extended in the published statistical literature. Furthermore, to our knowledge, no numerical procedure specifically designed to check for existence of the MLE has been developed yet. As a result, the possibility of the nonexistence of the MLE, even though well known, is rarely a concern for practitioners and is largely ignored, so that results and decisions stemming from the statistical analysis of tables containing zero counts are based on a possibly incorrect, faulty methodology. See the examples in Fienberg and Rinaldo (2006) and Example 6.5 below. Identifying the cases in which the MLE is not defined has immediate practical implications and is crucial for modifying traditional procedures of model selection based on both asymptotic and exact approximations of test statistics and, more generally, for developing new inferential methodologies to deal with sparse tables.

Recent advances in the field of algebraic statistics (Pistone et al., 2000; Diaconis and Sturm-

fels, 1998; Pachter and Sturmfels, 2005) have provided novel and broader mathematical tools for the analysis of categorical data. Their application has led to a series of theoretical results providing a complete, constructive characterization of the conditions for the existence of the MLE and proposing modifications to existing methods of inference for tables that do not have an MLE. These findings are contained in Eriksson et al. (2006), Rinaldo (2005) and Rinaldo (2006) and include (1) combinatorial and geometric characterization of the all possible patterns of sampling zeros leading to non-existence of the MLE; (2) design of a LP-based polynomial time algorithm for detecting non-existence; (3) theoretical derivation of the extended MLE and of its properties. Arising as a natural extension of the MLE, the extended MLE exhibits, in fact, all the defining features of the MLE and allows to identify "boundary" log-linear models for modelling sparse data that are informative only for some parameters of interest but not for all.

This document describes efficient and scalable computational strategies for extended maximum likelihood estimation, as described in Rinaldo (2006). The material presented here is taken, for the most part, from (Rinaldo, 2005, Chapter 6). We would like to acknowledge Stephen Fienberg, Mike Meyer and Pete Stewart for many of the results presented in Section 3.2 and 3.4 and for some of the pseudo codes of Section 7, which were derived from their unpublished work referenced as Fienberg et al. (1980).

This document is organized as follows. In Section 2 we give a technical background on maximum likelihood estimation for log-linear models for the Poisson and product-multinomial sampling schemes. Relevant results about existence of the MLE available in the statistical literature are reviewed and discussed. In particular, we summarize the most recent findings linking maximum likelihood estimation in log-linear models to extended exponential families and to polyhedral and algebraic geometry, contained in Rinaldo (2006). In Section 3 we offer a combinatorial, linear algebra and group-theoretical representation of the class of log-linear subspaces we are concerned in this work as a direct sum of orthogonal subspaces. We devise various ways of generating and manipulating the corresponding design matrices, which are computationally efficient and particularly suited for large problems. In Section 4, we derive a variety of numerical procedures for computing the extended MLE. Part of the section is devoted to the identification of those cells whose mean values cannot be determined by maximizing the likelihood because of insufficient information in the data. We also describe optimization procedure for the log-likelihood functions that are computationally efficient and take advantage of the algorithms from Section 3. We explore the connections between extended MLE and graphical models in Section 5. In particular, we will relate analytical factorization properties of the cell mean vectors with the combinatorial and graph-theoretical notions of reducibility and decomposability and we show how to take advantage of these features for the purpose of computing the extended MLE. In Section 6 we show by means of examples how to use the extended MLE for modifying, in a straightforward way, traditional χ^2 tests for goodness of fit and model selection. The pseudo codes for the algorithms we propose are in Section 7.

1.1 Notation

We introduce here the general, non-tabular notation for contingency tables and related quantities of interest that will be used throughout the document.

Consider K categorical random variables $X_1, \ldots, X_K$, each taking values on a finite set of labels, $\mathcal{I}_k = \{1, \ldots, I_k\}$, with $I_k \in \mathbb{N}_+$, $k = 1, \ldots, K$. Their cross-classification generates a set of label combinations, each called a *cell*, which is represented by the product set $\mathcal{I} = \bigotimes_{k=1}^{K} \mathcal{I}_k$. Every cell is uniquely identified by a K-dimensional multi-index $(i_1, \ldots, i_K) = \mathbf{i} \in \mathcal{I}$, whose k-th

coordinate indicates the value taken on by the k-th variable. To simplify the notation, the set of cells $\mathcal{I}$ will be represented as a lexicographically ordered linear list. This ordering is obtained through the bijection from $\mathcal{I}$ into $\left\{1, 2, \ldots, \prod_{k=1}^{K} I_k\right\}$ given by

$$< \mathbf{i} >=< i_1, \ldots, i_K > \rightarrow i_K + \sum_{k=1}^{K-1} \left(\prod_{j=k+1}^{K} I_j \right), \tag{1}$$

so that each K-tuple $\mathbf{i}$ will be unambiguously identified with its image $i =< \mathbf{i} >$ under the map (1). See Table 6 and Table 7 for the pseudo-codes implementing (1) and its inverse.

Example 1.1. Consider the 3-way table with levels $\mathcal{I}_j = \{1, 2, 3\}$, $1 \leq j \leq 3$. The table may be represented in tabular form as

111	121	131	211	221	231	311	321	331
112	122	132	212	222	232	312	322	332
113	123	133	213	223	233	313	323	333

and the corresponding numbering of the cells, according to Equation (1), is

1	4	7	2	5	8	3	6	9
10	13	16	11	14	17	12	15	18
19	22	25	20	23	26	21	24	27

. ■

Any set operation involving $\mathbf{i}$ will be expressed using the corresponding index i; for example, for $\mathcal{S} \subseteq \mathcal{I}$, $\mathbf{i} \in \mathcal{S}$ will be written $i \in \mathcal{S}$. Adopting this convention, $\mathcal{I}$ can be more conveniently thought of as the coordinate vector of $\mathbb{R}^{\mathcal{I}}$, the vector space of real-valued functions defined on $\mathcal{I}$. Then, the value of any $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ corresponding to the cell combination $\mathbf{i} \in \mathcal{I}$ will be indicated as $\mathbf{x}(i)$ or $\mathbf{x}_i$, where $i =< \mathbf{i} >$ is defined in (1). The standard inner product on $\mathbb{R}^{\mathcal{I}}$ is denoted by $(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{I}} x_i y_i$, with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$. If $s \subset \{1, \ldots, K\}$, then the coordinate projection of $\mathbf{i}$ onto $\mathcal{I}_s = \bigotimes_{k \in s} \mathcal{I}_k$ is the ordered list $\mathbf{i}_s = \{i_k \colon k \in s\}$, and will be written, using (1) again, as $i_s =< \mathbf{i}_s >$. The *s-margin* of $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ is the real-valued function on $\mathcal{I}_s$ with coordinate vector $\{i_s \colon i \in \mathcal{I}\} =< \mathcal{I}_s >$ obtained as

$$\mathbf{x}(i_s) = \sum_{j \in \mathcal{I} \colon j_s = i_s} \mathbf{x}(j),$$

and $\mathbf{x}(i_s)$ will be called the i_s-slice of $\mathbf{x}$. Similarly, for $\mathcal{S} \subset \mathcal{I}$, the restriction of $\mathbf{x}$ on $\mathcal{S}$ will be indicated with $\mathbf{x}_{\mathcal{S}}$. The support of a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ is defined to be the set $\{i \in \mathcal{I} \colon x_i \neq 0\}$ and is denoted $\mathrm{supp}(\mathbf{x})$. The set of vectors in $\mathbb{R}^{\mathcal{I}}$ with non-negative coordinates will be denoted by $\mathbb{R}^{\mathcal{I}}_{\geq 0}$. Matrices with coordinates indexed by subsets of $\mathcal{I}$ can be represented using the same convention and ordering. The column range of a matrix A will be denoted by $\mathcal{R}(\mathrm{A})$ and the orthogonal complement of a vector subspace $\mathcal{M}$ of $\mathbb{R}^{\mathcal{I}}$ will be written as $\mathcal{M}^{\perp}$. Functions and relations on vectors will be taken component-wise, unless otherwise specified. For example, for $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, $\exp^{\mathbf{x}} = \{\exp^{x_i} \colon i \in \mathcal{I}\}$ and $\mathbf{x} \geq \mathbf{0}$ means $x_i \geq 0$ for all i. The cardinality of a set B will be denoted by $|B|$.

2 Contingency Tables and Maximum Likelihood Estimation

Log-linear model analysis is concerned with the representation and study the joint distribution of the K variables $(X_1, \ldots, X_K)$. Data consists of a sequence of N independent realization, simultaneous realizations of the K factors and take the form of an unordered random sequence of labels $(L_1, \ldots, L_N)$, with $L_j \in \mathcal{I}$ for each $j = 1, \ldots, N$, where N itself can be random. A contingency table $\mathbf{n}$ is a sufficient statistic for the parameters of the underlying joint distribution of $(X_1, \ldots, X_K)$ in the form of cell counts.

Definition 2.1. A contingency table is a random function $\mathbf{n} \in \mathbb{R}^{\mathcal{I}}$ given by $\mathbf{n}(i) = \#\{j : L_j = i\}$.

The log-linear modeling approach hinges upon the representation of the cell mean vector $\mathbf{m} = \mathbb{E}[\mathbf{n}]$, which is assumed to be strictly positive, by means of a linear subspace $\mathcal{M}$ of $\mathbb{R}^{\mathcal{I}}$ containing $\boldsymbol{\mu} = \log \mathbf{m}$, to the extent that log-linear models themselves can in fact be defined by such subspaces. Namely, by fixing $\mathcal{M}$, it follows that the logarithms of the cell mean vectors must satisfy specific linear constraints, which completely characterize the underlying joint probability distribution. The defining subspace $\mathcal{M}$ will be called *log-linear subspace*.

The distribution of the cell counts is determined not only by the log-linear subspace, but also by the type of *sampling scheme* utilized in the collection of the data. A sampling scheme is a set of constraints on the observable cell counts induced by the sampling procedure. In this work, we will consider only sampling schemes defined by systems of linear forms. Formally, let $\mathcal{N} \subset \mathcal{M}$ be a linear subspace. The sampling restrictions dictated by $\mathcal{N}$ are of the form $(\boldsymbol{\gamma}_j, \mathbf{n}) = c_j$, $j = 1, \ldots, m$, where $(c_1, \ldots, c_m)$ are known constants and $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)$ are vectors spanning $\mathcal{N}$.

In particular, we will be focusing on the Poisson and product-multinomial sampling schemes, which are of widespread use. We remark that the log-linear framework allows for more general linear sampling designs, known as conditional Poisson schemes. Because a closed form expression for the probability mass function of these models is typically not available and inference is oftentimes infeasible, we chose not to treat them here. We refer to Haberman (1974) and Rinaldo (2006) for background and results on these general models. We also point out that our results can be applied in a straightforward way to the mixed Poisson-multinomial models described by Lang (2004).

2.1 Sampling Schemes and Log-likelihood functions

2.1.1 Poisson Sampling Scheme

Under the Poisson sampling scheme, $\mathcal{N}$ is the trivial subspace $\{\mathbf{0}\}$ and the sampling is unconstrained. Note that the number of data points N must then be random. As a result, the components of $\mathbf{n}$ are independent Poisson random variables with $\mathbb{E}[n_i] = m_i$, for $i \in \mathcal{I}$. The log-likelihood function $\ell_{\mathcal{P}}$, as a function of $\boldsymbol{\mu} \in \mathcal{M}$, is

$$\ell_{\mathcal{P}}(\boldsymbol{\mu}) = (\mathbf{n}, \boldsymbol{\mu}) - \sum_{i \in \mathcal{I}} \exp^{\mu_i} - \sum_{i \in \mathcal{I}} \log n_i!. \tag{2}$$

2.1.2 Product Multinomial Sampling Scheme

Under the product-multinomial sampling scheme, the observed vector of counts consists of one or more independent multinomial random vectors. Let

$$\mathcal{I} = \biguplus_{j=1}^{r} \mathcal{B}_j$$

be a partition of the set $\mathcal{I}$ into r classes, where $\uplus$ denotes disjoint union. For each partitioning set $\mathcal{B}_j$ let $\boldsymbol{\chi}_j \in \mathbb{R}^{\mathcal{I}}$ be its indicator function, which is given by

$$\boldsymbol{\chi}_j(i) = \begin{cases} 1 & \text{if } \ i \in \mathcal{B}_j \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and define $\mathcal{N}$ to be the r-dimensional subspace spanned by the orthogonal vectors $(\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r)$. The product-multinomial sampling constraints require that $(\mathbf{n}, \boldsymbol{\chi}_j) = (\mathbf{m}, \boldsymbol{\chi}_j) = N_j$ for fixed positive integers N_j so that the joint distribution of the cell counts is the product of r independent multinomial distributions with sizes N_j, each supported on the set $\mathcal{B}_j$.

Typically, the spanning vectors of $\mathcal{N}$ are defined in a more intuitive way using indicator functions of slices of $\mathbb{R}^{\mathcal{I}}$. Specifically, let $b \subset \{1, \ldots, K\}$ and $\mathcal{I}_b = \bigotimes_{k \in b} \mathcal{I}_k$ and, for each $j \in \mathcal{I}_b$, define $\mathcal{B}_j = \{i \in \mathcal{I} \colon i_b = j\}$. By construction $\mathcal{I} = \biguplus_{\mathbf{j} \in \mathcal{I}_b} \mathcal{B}_j$ and $\mathcal{N}$ is the r-dimensional subspace spanned by the orthogonal vectors $\{\boldsymbol{\chi}_j\}_{j \in \mathcal{I}_b}$, where $\boldsymbol{\chi}_j$ are defined as in (3) and $r = |\mathcal{I}_b|$. For some authors, e.g. Lauritzen (1996), this is in fact taken to be the definition of the product-multinomial scheme.

The log-likelihood function at a point $\boldsymbol{\mu} \in \mathcal{M}$ is (see Haberman, 1974, Equation 1.51)

$$\ell_{\mathcal{L}}(\boldsymbol{\mu}) = \sum_{j=1}^{r} \left(\sum_{i \in \mathcal{B}_j} n_i \log \frac{m_i}{(\mathbf{m}, \boldsymbol{\chi}_j)} + \log N_j! - \sum_{i \in \mathcal{B}_j} n_i! \right), \tag{4}$$

where the cell mean vector is $\mathbf{m} = \exp^{\boldsymbol{\mu}}$. For the trivial partition with only one class, there is only one constraint, namely $(\mathbf{1}_{\mathcal{I}}, \mathbf{n}) = N$, and the distribution of the cell counts is multinomial with size N. Because of the sampling constraints, the log-likelihood (4) is well defined only on the subset $\widetilde{\mathcal{M}}$ of $\mathcal{M}$ given by

$$\widetilde{\mathcal{M}} = \{\boldsymbol{\mu} \in \mathcal{M} \colon (\boldsymbol{\chi}_j, \exp^{\boldsymbol{\mu}}) = N_j, j = 1, \ldots, r\}. \tag{5}$$

Note that $\widetilde{\mathcal{M}}$ is neither a vector space nor a convex set. A more convenient parametrization for $\ell_{\mathcal{L}}$ leading to a more manageable parameter space can be obtained as follows. Let $\mathcal{M} \ominus \mathcal{N} = \mathcal{M} \cap \mathcal{N}^{\perp}$.

Lemma 2.2. *The following injective function on $\mathcal{M} \ominus \mathcal{N}$*

$$\ell_{\mathcal{L}}(\boldsymbol{\beta}) = (\mathbf{n}, \boldsymbol{\beta}) - \sum_{j=1}^{r} N_j \log(\exp^{\boldsymbol{\beta}}, \boldsymbol{\chi}_j) - \sum_{i \in \mathcal{I}} n_i!, \quad \forall \boldsymbol{\beta} \in \mathcal{M} \ominus \mathcal{N}. \tag{6}$$

parametrizes the log-likelihood function for the product-multinomial model.

This re-parametrization is essentially equivalent to reduction to minimal form of the underlying exponential family of distributions for the cell counts via sufficiency (Theorem 3.5 in Rinaldo, 2006) and offers considerable computational advantages, which will be exploited in Section 4.

Proof. We first show that there is a one-to-one correspondence between $\mathcal{M} \ominus \mathcal{N}$ and $\widetilde{\mathcal{M}}$. It is clear that to any point $\tilde{\boldsymbol{\mu}}$ there corresponds a unique point $\tilde{\boldsymbol{\beta}} = \Pi_{\mathcal{M}\ominus\mathcal{N}}\tilde{\boldsymbol{\mu}} \in \mathcal{M} - \mathcal{N}$ and a unique point $\tilde{\boldsymbol{\nu}} = \Pi_{\mathcal{N}}\tilde{\boldsymbol{\mu}} \in \mathcal{N}$ such that $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\nu}}$. To show the converse, let $\beta \in \mathcal{M} \ominus \mathcal{N}$. Next, notice that if $\boldsymbol{\nu} \in \mathcal{N}$, then $\boldsymbol{\nu} = \sum_{j=1}^{r} c_j \boldsymbol{\chi}_j$ for some constants $c_1, \ldots, c_r$ and, because $\operatorname{supp}(\boldsymbol{\chi}_j) \cap \operatorname{supp}(\boldsymbol{\chi}_k) = \emptyset$, for $j \neq k$, it must be that $\boldsymbol{\nu}(i) = c_j$, for all $i \in \mathcal{B}_j$. Then,

$$\exp^{c_j}(\exp^{\boldsymbol{\beta}}, \boldsymbol{\chi}_j) = \exp^{c_j} \sum_{i \in \mathcal{B}_j} \exp^{\beta_i} = (\exp^{\boldsymbol{\beta}+\boldsymbol{\nu}}, \boldsymbol{\chi}_j). \tag{7}$$

For each j, let $\hat{c}_j = \log\left(\frac{N_j}{(\exp^{\boldsymbol{\beta}}, \boldsymbol{\chi}_j)}\right)$ (notice that these coefficients depend on $\boldsymbol{\beta}$ only) and let $\boldsymbol{\nu}_{\boldsymbol{\beta}} = \sum_{j=1}^{r} \hat{c}_j \boldsymbol{\chi}_j \in \mathcal{N}$. Then, equation (7) implies that $\boldsymbol{\beta} + \boldsymbol{\nu}_{\boldsymbol{\beta}} \in \widetilde{\mathcal{M}}$. Therefore, for each $\tilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}$, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\beta} + \boldsymbol{\nu}_{\boldsymbol{\beta}}$, where $\boldsymbol{\nu}_{\boldsymbol{\beta}}$ is obtained by equating (7) to N_j. Hence,

$$(\tilde{\boldsymbol{\mu}}, \mathbf{n}) = (\mathbf{n}, \boldsymbol{\beta}) - \sum_{j=1}^{k} N_j \log(\exp^{\boldsymbol{\beta}}, \boldsymbol{\chi}_j) + \sum_{j=1}^{r} N_j \log N_j.$$

Using (4), direct calculation shows that

$$\ell_{\mathcal{L}}(\tilde{\boldsymbol{\mu}}) = (\mathbf{n}, \tilde{\boldsymbol{\mu}}) - \sum_{j=1}^{r} N_j \log N_j - \sum_{i \in \mathcal{I}} n_i! = \ell_{\mathcal{L}}(\boldsymbol{\beta}),$$

and the proof is complete. ■

2.2 Existence of the MLE.

Consider, for convenience, the parametrization of the log-likelihood functions $\ell_{\mathcal{P}}$ and $\ell_{\mathcal{L}}$ in terms of points of $\mathcal{M}$ and $\widetilde{\mathcal{M}}$, respectively. The MLE of $\boldsymbol{\mu} = \log \mathbf{m}$ is the set

$$\{\boldsymbol{\mu}^* \in \mathcal{M} \colon \ell_{\mathcal{P}}(\boldsymbol{\mu}^*) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \ell_{\mathcal{P}}(\boldsymbol{\mu})\}$$

for the Poisson likelihood and the set

$$\{\boldsymbol{\mu}^* \in \widetilde{\mathcal{M}} \colon \ell_{\mathcal{M}}(\boldsymbol{\mu}^*) = \sup_{\tilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}} \ell_{\mathcal{M}}(\tilde{\boldsymbol{\mu}})\}$$

for the product-multinomial likelihood. The MLE is said to be nonexistent if the supremum is not attained at any point in the appropriate log-linear sub-space, i.e. if the above set is empty. With a slight abuse of language, since, in the present problem, the MLE is shown to be always a single point rather than a set, we will always speak of the MLE as a vector rather than a set. Note that, under both sampling schemes, when the MLE of $\boldsymbol{\mu}$ exists, the MLE of $\mathbf{m}$ is a strictly positive vector. As noted in Haberman (1974), both the log-likelihood functions $\ell_{\mathcal{P}}$ and $\ell_{\mathcal{M}}$ are concave and bounded from above, so that the nonexistence of the MLE is caused by the behavior of the likelihoods at the boundary of the parameter space. In fact, the main result described later is that nonexistence of the MLE can be characterized by sequences of points in the domain of the log-likelihood function that realize the supremum in the limit but, at the same time, have norms exploding to infinity.

The directions of recession along any of such sequences will correspond to non-estimable natural parameters and non-estimable cell mean counts. As a result, when the MLE of $\mathbf{m}$ does not exist, the optimization of the log-likelihoods will lead to an estimated cell mean vector with some zero coordinates. The possibility of identifying the coordinates that correspond to the non-estimable parameters given the data, along with their statistical interpretation, is one the contributions in this work.

2.2.1 Literature Review

We give here a brief history of the main contributions to the theory of maximum likelihood estimation in log-linear models. For a more detailed account see Fienberg and Rinaldo (2006).

Birch (1963) conducted the first rigorous study of the conditions for existence of the maximum likelihood estimate of the cell mean vector in log-linear models. The author considered hierarchical log-linear models and showed that, under the assumption $\mathbf{n} > \mathbf{0}$, the maximum likelihood estimate of $\mathbf{m}$ exists uniquely and satisfies $\mathrm{P}_{\mathcal{M}}\mathbf{n} = \mathrm{P}_{\mathcal{M}}\mathbf{m}$, where $\mathrm{P}_{\mathcal{M}}$ is the projection matrix onto the log-linear subspace $\mathcal{M}$ (although the author did not formalize his result in this fashion). In addition, Birch (1963) showed that, if $\mathrm{P}_{\mathcal{N}}\mathbf{n} = \mathrm{P}_{\mathcal{N}}\mathbf{m}$, the MLE is the same for both Poisson and multinomial sampling schemes. Birch (1963)'s findings were greatly generalized by Haberman (1974), whose results have represented the most thorough and advanced treatment on this subject for many years and are reproduced for completeness below, without proof.

Haberman (1974), Theorem 2.1.
Under Poisson scheme, if a maximum likelihood estimate $\widehat{\boldsymbol{\mu}}$ exists, then it is unique and satisfies

$$\mathrm{P}_{\mathcal{M}}\mathbf{n} = \mathrm{P}_{\mathcal{M}}\widehat{\mathbf{m}}, \tag{8}$$

where $\widehat{\mathbf{m}} = \exp\widehat{\boldsymbol{\mu}}$. Conversely, if for some $\widehat{\boldsymbol{\mu}} \in \mathcal{M}$ and $\widehat{\mathbf{m}} = \exp\widehat{\boldsymbol{\mu}}$ Equation (8) is satisfied, then $\widehat{\mathbf{m}}$ is the maximum likelihood estimate of $\mathbf{m}$.

Haberman (1974), Theorem 2.2.
Under Poisson scheme, the maximum likelihood estimate of $\mathbf{m}$ exists if and only if there exists a $\boldsymbol{\delta} \in \mathcal{M}^{\perp}$ such that $\mathbf{n} + \boldsymbol{\delta} > \mathbf{0}$.

Haberman (1974), Theorem 2.3.
Under Poisson scheme, the maximum likelihood estimate of $\mathbf{m}$ exists if and only if there does not exist any $\boldsymbol{\mu} \in \mathcal{M}$ such that $\boldsymbol{\mu} \gneq \mathbf{0}$ and $(\mathbf{n}, \boldsymbol{\mu}) = 0$.

Haberman (1974), Theorem 2.4.
Provided, $\mathcal{N} \subset \mathcal{M}$, the maximum likelihood estimate of $\mathbf{m}$ under Poisson sampling scheme exists if and only if it exists under product-multinomial sampling scheme. If they exist, they coincide.

Theorem 2.2 in Haberman (1974) formalizes the intuition that the MLE is defined whenever it is possible to "eliminate" the zero cells in the table by adding and subtracting appropriate quantities to the observed table $\mathbf{n}$. These results not only explain why the MLE is non-existent if one of the margins is null but also allowed Haberman to give an example of a table for which the MLE is not defined and the margins are positive, describe in Example 6.1, a case termed "pathological". Despite the concerned raised by some authors (see Bishop et al., 1975) about the potential deleterious

effect of pathological configurations of zeros in large and sparse tables, Example 6.1 has been, up until very recently, the only published example of a contingency table with such feature. This was largely due to the fact that, though very intuitive, Haberman's characterization is non-constructive and ultimately impractical, in the sense that it does not lead to efficient algorithms for checking the existence of the MLE.

Finally, Haberman (1974, Appendix B) noted by means of examples that, when the MLE is not defined, the maximization of the log-likelihood function produce a sequence of points converging to a unique maximizer $\widehat{\mathbf{m}}^{\mathrm{e}}$, some of whose coordinates, typically a subset of the zero cells, are 0. Such estimate was then heuristically called *extended MLE*.

In an unpublished manuscript, Fienberg et al. (1980) offered a different derivation of Haberman (1974)'s results for the Poisson and product-multinomial sampling schemes and provided a formal definition of the extended MLE. The main goal in their work was the development of efficient algorithms for identifying cases of nonexistent MLE and computing the extended MLE. Their approach can be described as follows. For a given observed table $\mathbf{n}$, let $\mathcal{I}_+ := \mathrm{supp}(\mathbf{n})$ and $\mathcal{I}_0 := \mathcal{I} \backslash \mathrm{supp}(\mathbf{n})$ be the set of cells with positive and zero counts, respectively. Define the *critical set* $\mathcal{C}$ to be be the subset of $\mathcal{I}_0$ with maximal cardinality such that there exists a $\boldsymbol{\mu} \in \mathcal{M}$ with $\boldsymbol{\mu}_{\mathcal{C}} < \mathbf{0}$ but $\boldsymbol{\mu}_{\mathcal{C}^c} = \mathbf{0}$, where $\mathcal{C}^c = \mathcal{I}_0 \backslash \mathcal{C}$. Vectors satisfying these conditions are called *critical vectors*. Next, let $\overline{M}$ denote the sequential closure in $\mathbb{R}^{\mathcal{I}}$ of the set $\{\exp^{\boldsymbol{\mu}} \colon \boldsymbol{\mu} \in \mathcal{M}\}$.

Theorem 2.3 (Fienberg et al. (1980))**.** *The maximum likelihood estimate $\widehat{\mathbf{m}}$ of $\mathbf{m}$ exists if and only if $\mathcal{C} = \emptyset$. If $\mathcal{C} \neq \emptyset$, then the unique extended maximum likelihood estimate $\widehat{\mathbf{m}}^{\mathrm{e}}$, identical for both the Poisson and product-multinomial sampling scheme, exists and satisfies these two defining conditions:*

1. $\widehat{\mathbf{m}}^{\mathrm{e}}_{\mathcal{C}} = \mathbf{0}$,
2. $\widehat{\mathbf{m}}^{\mathrm{e}}$ *is the only vector in* $\overline{M}$ *such that* $\mathrm{P}_{\mathcal{M}}\mathbf{n} = \mathrm{P}_{\mathcal{M}}\widehat{\mathbf{m}}^{\mathrm{e}}$.

Note that the condition for the existence of the MLE is a restatement of Haberman's Theorem 2.2. In fact, if $\mathcal{C} \neq \emptyset$, then there exists $\boldsymbol{\mu} \in \mathcal{M}$ such that $(\mathbf{n}, \boldsymbol{\mu}) = 0$, $\boldsymbol{\mu} \lneq 0$. On the other hand, if there exists a $\boldsymbol{\mu} \in \mathcal{M}$ such that $(\mathbf{n}, \boldsymbol{\mu}) = 0$, $\boldsymbol{\mu} \lneq \mathbf{0}$, then $\mathcal{C} \supseteq \{i \in \mathcal{I} \colon \mu_i < 0\} \neq \emptyset$. The novelty in the contribution of Fienberg et al. (1980) lies in the characterization of the critical set and in the consequent formal definition of the extended MLE as the unique maximizer of the log-likelihood function. Furthermore, and quite importantly, the extended MLE is shown to be always defined, no matter how sparse a table is, and to possess properties mirroring the ones of the "ordinary" MLE.

Gloneck et al. (1988) proved, by means of counter-examples, that positivity of the margins is a necessary and sufficient conditions for existence of the MLE if and only if the model is decomposable.

Lauritzen (1996) used a slightly different approach to maximum likelihood estimation in hierarchical log-linear models. In fact, rather than the log-linear subspace $\mathcal{M}$, its sequential closure $\overline{\mathcal{M}}$ is taken to be the appropriate parameter space and hence the domain of the Poisson and product-multinomial log-likelihood functions (2) and (4). Because of this enlarged parameter space, Lauritzen (1996) referred to log-linear models as *extended log-linear models*. In Theorem 4.8 and Theorem 4.11 the author summarized this characterization of the MLE, which is, as far as existence and uniqueness are concerned, virtually identical to the one for extended MLE offered by Fienberg et al. (1980). However, Lauritzen (1996)'s results remain overall non-constructive and of limited practical use. Furthermore the author does not make an explicit distinction between MLE and extended MLE is made, and does not offer any interpretation of the MLE obtained in an extended sense.

2.2.2 Recent Results

In this section we summarize the latest results on maximum likelihood estimation for log-linear models. Both the methodology and the results were largely inspired by recent advances in the field of algebraic statistics (Pistone et al., 2000; Diaconis and Sturmfels, 1998; Pachter and Sturmfels, 2005), which have indicated a more general approach to the study of log-linear models that takes advantage of the tolls and formalism of algebraic and polyhedral geometry.

Eriksson et al. (2006) used Theorem 2.1 in Haberman (1974) to provide geometric conditions for the existence of the MLE in Poisson and product-multinomial schemes for hierarchical log-linear models. Their characterization is relatively simple, as it relies on basic results form polyhedral geometry, but, at the same time, proved itself quite powerful and suited for numerical implementation. In order to describe their results, basic notions from polyhedral geometry will be introduced here, which will be used throughout the paper. See, in particular, Ziegler (1998). For a given log-linear subspace $\mathcal{M}$, let A be the corresponding $d \times |\mathcal{I}|$ design matrix whose row span is $\mathcal{M}$. We don't require A to be of full rank, i.e. d may be bigger than $\mathrm{rank}(\mathrm{A})$, and we will always assume that the entries of A are integral, a hypothesis that is hardly restrictive and naturally satisfied by the log-linear subspaces we study in Section 3. Consider the polyhedral cone generated by the columns of the matrix A,

$$\mathrm{C_A} = \{\boldsymbol{\xi} \colon \boldsymbol{\xi} = \mathrm{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}\},$$

which is called the *marginal cone*. Eriksson et al. (2006) showed that, provided that $\mathcal{N} \subset \mathcal{M}$, the MLE of the cell mean vector under both Poisson and product-multinomial schemes exists if and only if $\mathrm{A}\mathbf{n} \in \mathrm{ri}(\mathrm{C_A})$. The d-dimensional integer vector $\mathbf{t} = \mathrm{A}\mathbf{n}$ is minimal sufficient statistic for the vector of parameters $\boldsymbol{\mu}$, so the condition for the existence of the MLE reduces to the study of the geometric properties of the sufficient statistics. This result was further generalized by Rinaldo (2006) who studied maximum likelihood estimation for log-linear models using the theory of exponential families, polyhedral and algebraic geometry. The remainder of the section summarizes these findings.

A face of $\mathrm{C_A}$ is a set $F = \{\boldsymbol{\xi} \in \mathrm{C_A} \colon (\boldsymbol{\xi}, \zeta_F) = 0\}$, for some $\zeta_F \in \mathbb{R}^d$ such that $(\boldsymbol{\xi}, \zeta_F) \leq 0$ for all $\boldsymbol{\xi} \in \mathrm{C_A}$. Note that the polyhedral cone $\mathrm{C_A}$ has a finite number of faces and that $\mathrm{C_A}$ is a face of itself, termed improper. To every face F of $\mathrm{C_A}$ there corresponds a subset $\mathcal{F}$ of $\mathcal{I}$ such that $(\mathbf{a}_i, \boldsymbol{\zeta}_F) = 0$, for all $i \in \mathcal{F}$ and $(\mathbf{a}_i, \boldsymbol{\zeta}_F) < 0$ for all $i \in \mathcal{F}^c = \mathcal{I} \setminus \mathcal{F}$. In words, the set $\mathcal{F}$ consists of the cell indexes of the columns of A whose conic hull is precisely F. Conversely, if there exists a vector $\boldsymbol{\zeta}$ that defines a set $\mathcal{F} \subset \mathcal{I}$ with the above properties, then $\boldsymbol{\zeta}$ determines the supporting hyperplane $H_F = \{\mathbf{x} \in \mathbb{R}^d \colon (\boldsymbol{\zeta}, \mathbf{x}) = 0\}$ for F, i.e. $F = \mathrm{C_A} \cap H_F$. The set $\mathcal{F}$ is called the facial set of F (Geiger et al., 2006). For any $\boldsymbol{\xi} \in \mathrm{C_A}$ there exists only one (possibly improper) face F containing $\boldsymbol{\xi}$ in its relative interior, i.e. such that $\mathbf{t}$ is a linear combination with positive coefficients of the columns of A with indexes in $\mathcal{F}$. This implies that, for any $\boldsymbol{\xi} \in \mathrm{C_A}$, there is one facial set $\mathcal{F}$ such that $\boldsymbol{\xi} = \mathrm{A}\mathbf{x}$, for some $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ with $\mathrm{supp}(\mathbf{x}) = \mathcal{F}$. The facial sets provide a combinatorial representation of the face lattice of $\mathrm{C_A}$ and play a fundamental role in identifying the set of cells for which an MLE of the expected count is not defined. In particular, for a given table $\mathbf{n}$, we will be concerned with determining the (random) facial set corresponding to the sufficient statistics $\mathbf{t} = \mathrm{A}\mathbf{n}$ (see Section 4.1).

The next result is a generalization of Theorem Fienberg et al. (1980) showing that critical sets are in fact facial sets. The proof is a simplified version of arguments described in detail in Rinaldo (2006).

Theorem 2.4. *Assume $\mathcal{N} \subset \mathcal{M}$ and let $\mathcal{F}$ be the facial set associated to $\mathbf{t}$. The MLE $\widehat{\mathbf{m}}$ of the cell mean vector exists, is unique and identical under Poisson and product-multinomial if and only if $\mathcal{F} = \mathcal{I}$. If $\mathcal{F} \subsetneq \mathcal{I}$, there exists one point $\widehat{\mathbf{m}}^{\mathrm{e}}$ in $\overline{M}$ such that $\widehat{\mathbf{m}}^{\mathrm{e}} = \lim_n \exp\{\boldsymbol{\mu}_n\}$, where $\{\boldsymbol{\mu}\}_n \subset \mathcal{M}$ is a sequence for which $\lim_n \ell_{\mathcal{P}}(\boldsymbol{\mu}_n) = \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{P}}(\boldsymbol{\mu})$ and $\lim_n \ell_{\mathcal{L}}(\boldsymbol{\mu}_n) = \sup_{\tilde{\boldsymbol{\mu}}\in\widetilde{\mathcal{M}}} \ell_{\mathcal{L}}(\tilde{\boldsymbol{\mu}})$. Furthermore, $\mathrm{supp}(\widehat{\mathbf{m}}^{\mathrm{e}}) = \mathcal{F}$ and $\mathrm{P}_{\mathcal{M}}\mathbf{n} = \mathrm{P}_{\mathcal{M}}\widehat{\mathbf{m}}^{\mathrm{e}}$.*

Proof. We show that, under both Poisson and product multinomial scheme, the MLE exists unique and is identical in both cases if and only if $\mathbf{t} = \mathrm{A}\mathbf{n}$ is a point in the relative interior of $\mathrm{C_A}$. If $\mathbf{t}$ belongs to the relative interior of a face F, then both the log-likelihoods realize their suprema along sequences of points $\boldsymbol{\mu}_n \subset \mathcal{M}$ for which the limit $\exp^{\boldsymbol{\mu}_n} = \widehat{\mathbf{m}}$ is unique, satisfies the moment equations $\mathrm{P}_{\mathcal{M}}\mathbf{n} = \mathrm{P}_{\mathcal{M}}\widehat{\mathbf{n}}$ and $\mathrm{supp}(\widehat{\mathbf{m}}) = \mathcal{F}$.

First, we consider the problem of maximizing the log-likelihood $\ell_{\mathcal{P}}(\mu) = (\mathbf{n}, \mu) - \sum_{i\in\mathcal{I}} \exp^{\mu_i}$ under Poisson sampling scheme. Suppose $\mathbf{t} = \mathrm{A}\mathbf{n}$ lies inside the relative interior of a proper face F of $\mathrm{C_A}$. Then, there exists a $\mathbf{z}_F \in \mathrm{kernel}(\mathrm{A})$ such that the vector $\mathbf{x}_F = \mathbf{n} + \mathbf{z}_F$ satisfies $\mathbf{t} = \mathrm{A}\mathbf{x}_F$ and $\mathrm{supp}(\mathbf{n} + \mathbf{z}_F)$ gives the corresponding facial set $\mathcal{F}$. Then, $\ell_{\mathcal{P}}(\mu) = (\mathbf{x}_F, \mu) - \sum_{i\in\mathcal{I}} \exp^{\mu_i}$, since, for $\mu \in \mathcal{M}$, $(\mathbf{z}_F, \mu) = 0$.

Let $\pi_{\mathcal{F}} \colon \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{F}}$ and $\pi_{\mathcal{F}^c} \colon \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{F}^c}$ be the coordinate projection maps from $\mathcal{I}$ into $\mathcal{F}$ and $\mathcal{F}^c$, respectively. Define $\ell_{\mathcal{F}}$ and $\ell_{\mathcal{F}^c}$ to be the restriction of $\ell_{\mathcal{P}}$ on $\pi_{\mathcal{F}}(\mathcal{M})$ and $\pi_{\mathcal{F}^c}(\mathcal{M})$, respectively. Explicitly, $\ell_{\mathcal{F}}(\boldsymbol{\mu}) = (\mathbf{x}_F, \pi_{\mathcal{F}}(\mu)) - \sum_{i\in\mathcal{F}} \exp^{\mu_i} = (\mathbf{x}_F, \mu) - \sum_{i\in\mathcal{F}} \exp^{\mu_i}$ and $\ell_{\mathcal{F}^c}(\boldsymbol{\mu}) = -\sum_{i\in\mathcal{F}^c} \exp^{\mu_i}$. Therefore, $\ell_{\mathcal{P}}(\boldsymbol{\mu}) = \ell_{\mathcal{F}}(\boldsymbol{\mu}) + \ell_{\mathcal{F}^c}(\boldsymbol{\mu})$. The function $\ell_{\mathcal{F}}$ is continuous and strictly concave on $\pi_{\mathcal{F}}(\mathcal{M})$ and is bounded from above, since $\lim_{\boldsymbol{\mu}\colon ||\pi_{\mathcal{F}}(\boldsymbol{\mu})||\to\infty} \ell_{\mathcal{F}}(\boldsymbol{\mu}) = -\infty$. Therefore, $\ell_{\mathcal{F}}$ achieves its supremum at a point in $\pi_{\mathcal{F}}(\mathcal{M})$ with finite coordinates. The function $\ell_{\mathcal{F}^c}$ is negative and strictly decreasing in each coordinate of its argument and $\sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}^c}(\boldsymbol{\mu}) = 0$. Conclude that $\sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{P}}(\boldsymbol{\mu}) = \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}}(\boldsymbol{\mu})$.

Pick any sequence $\{\boldsymbol{\gamma}_n\}_n \subset \mathcal{M}$ for which $\lim_n \ell_{\mathcal{F}}(\boldsymbol{\gamma}_n) = \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}}(\boldsymbol{\mu})$ (note that necessarily $\pi_{\mathcal{F}}(\boldsymbol{\gamma}) = \lim_n \pi_{\mathcal{F}}(\boldsymbol{\gamma}_n)$ belongs to $\pi_{\mathcal{F}}(\mathcal{M})$ and has finite norm.) Next, choose a sequence $\{\boldsymbol{\nu}_n\}_n \subset \mathcal{M}$ such that $\lim_n \frac{\boldsymbol{\gamma}_n(i)}{\boldsymbol{\nu}_n(i)} = 0$ for all $i \notin \mathcal{F}$ and $\lim_n \boldsymbol{\nu}_n(i) = -\infty$ for all $i \in \mathcal{F}$ and $\boldsymbol{\nu}_n(i) = 0$ for all n and $i \notin \mathcal{F}$. Since $\mathcal{F}$ is a facial set, sequences with these properties exist. Then,

$$\lim_n \ell_{\mathcal{F}^c}(\boldsymbol{\nu}_n + \boldsymbol{\gamma}_n) = \lim_n \ell_{\mathcal{F}^c}(\boldsymbol{\nu}_n) = 0 = \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}^c}(\boldsymbol{\mu}).$$

Consider now the new sequence $\{\boldsymbol{\mu}_n\}_n \subset \mathcal{M}$, where $\boldsymbol{\mu}_n = \boldsymbol{\gamma}_n + \boldsymbol{\nu}_n$. Then

$$\begin{aligned}
\lim_n \ell_{\mathcal{P}}(\boldsymbol{\mu}_n) &= \lim_n \left(\ell_{\mathcal{F}}(\boldsymbol{\gamma}_n + \boldsymbol{\nu}_n) + \ell_{\mathcal{F}^c}(\boldsymbol{\gamma}_n + \boldsymbol{\nu}_n)\right) \\
&= \lim_n \ell_{\mathcal{F}}(\boldsymbol{\gamma}_n) + \lim_n \ell_{\mathcal{F}^c}(\boldsymbol{\nu}_n) \\
&= \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}}(\boldsymbol{\mu}) + \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}^c}(\boldsymbol{\mu}) \\
&= \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{F}}(\boldsymbol{\mu}) \\
&= \sup_{\boldsymbol{\mu}\in\mathcal{M}} \ell_{\mathcal{P}}(\boldsymbol{\mu})
\end{aligned}$$

Let $\widehat{\mathbf{m}} = \lim_n \exp^{\boldsymbol{\mu}_n}$. Then $\widehat{\mathbf{m}}$ is a non-negative vector with support $\mathcal{F}$ such that $\pi_{\mathcal{F}}(\widehat{\mathbf{m}}) = \exp\{\pi_{\mathcal{F}}(\boldsymbol{\gamma})\}$. Furthermore, since $\ell_{\mathcal{F}}$ admits a unique maximizer, the optimum $\widehat{\mathbf{m}}$ must be unique.

Next, since $\boldsymbol{\gamma} = \lim_n \boldsymbol{\gamma}_n$ maximizes $\ell_{\mathcal{F}}$, the first order conditions on the differential of $\ell_{\mathcal{F}}$ (see Haberman, 1974, Chapter 2) gives

$$(\pi_{\mathcal{F}}(\boldsymbol{\lambda}), \exp\{\pi_{\mathcal{F}}(\boldsymbol{\gamma})\}) = (\pi_{\mathcal{F}}(\boldsymbol{\lambda}), \pi_{\mathcal{F}}(\mathbf{x}_F)) = (\boldsymbol{\lambda}, \mathbf{n}),$$

for all $\boldsymbol{\lambda} \in \mathcal{M}$, where the last equality stems from the fact that $\pi_{\mathcal{F}^c}(\mathbf{x}_F) = \mathbf{0}$. But this, in turn, implies that

$$\left(\boldsymbol{\lambda}, \exp^{\boldsymbol{\mu}^*}\right) = (\boldsymbol{\lambda}, \mathbf{n})$$

for all $\boldsymbol{\lambda} \in \mathcal{M}$, where $\boldsymbol{\mu}^* = \lim_n \boldsymbol{\mu}_n$, and hence

$$\mathrm{P}_{\mathcal{M}} \widehat{\mathbf{m}} = \mathrm{P}_{\mathcal{M}} \mathbf{n}. \tag{9}$$

If instead the log-likelihood function $\ell_{\mathcal{L}}$ under product-multinomial sampling is to be maximized, it is necessary to consider only the points $\tilde{\boldsymbol{\mu}}$ inside $\widetilde{\mathcal{M}}$ as in Equation (5). Fortunately, this restriction is inconsequential. In fact, first note that, by (9) and because $\mathcal{N} \subset \mathcal{M}$, the limit $\boldsymbol{\mu}^*$ satisfies the constraints $\{(\boldsymbol{\chi}_j, \exp^{\boldsymbol{\mu}}) = N_j, j = 1, \ldots, r\}$. Next, since $\ell_{\mathcal{L}}$ and $\ell_{\mathcal{P}}$ differ by a constant on $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}} \subset \mathcal{M}$, we have that

$$\ell_{\mathcal{L}}(\boldsymbol{\mu}^*) = \sup_{\tilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}} \ell_{\mathcal{L}}(\tilde{\boldsymbol{\mu}}).$$

Conclude that the log-likelihood functions under both the Poisson and product multinomial model have the same maximizer $\widehat{\mathbf{m}}$.

Finally, notice that if $\mathbf{t} \in \mathrm{ri}(\mathrm{C_A})$, so that $\mathcal{F} = \mathcal{I}$, the arguments simplify. Explicitly, there exists a point $\boldsymbol{\mu}^* \in \widetilde{\mathcal{M}} \subset \mathcal{M}$ such that

$$\begin{array}{rcl} \sup_{\boldsymbol{\mu} \in \mathcal{M}} \ell_{\mathcal{P}}(\boldsymbol{\mu}) & = & \ell_{\mathcal{P}}(\boldsymbol{\mu}^*) \\ \sup_{\tilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}} \ell_{\mathcal{L}}(\tilde{\boldsymbol{\mu}}) & = & \ell_{\mathcal{L}}(\boldsymbol{\mu}^*) \end{array} .$$

■

The previous theorem shows that, for any observed table $\mathbf{n}$, one maximizer $\widehat{\mathbf{m}}^{\mathrm{e}}$ of the log-likelihood functions can always be found and it exhibits the same features as the "ordinary" MLE.

Definition 2.5. The vector $\widehat{\mathbf{m}}^{\mathrm{e}}$ from Theorem 2.4 is the *extended MLE* of $\mathbf{m}$.

In fact, $\widehat{\mathbf{m}}^{\mathrm{e}}$ is the MLE of the cell mean vector for a "boundary" log-linear model in the closure parameter space, under mean-value parametrization. More generally, extended MLEs arise as regular MLEs for an the extended exponential family of distributions supported on the facial sets of $\mathrm{C_A}$ and that the MLE is just a special case of extended MLE induced by observed sufficient statistics in the interior of the marginal cone rather than on its boundary. The additional distributions making up for the extended exponential family have cell mean vectors supported on the facial sets and are derived by closing appropriately the parameter space, as described below.

The results stated in Theorem 2.4 are derived using only analytic arguments focusing on the maximization of the log-likelihood function. Rinaldo (2006) gives a different derivation, based on further geometric properties of log-linear models. This approach, which relies on notion from algebraic geometry (see, e.g., Cox et al., 1996), is particularly convenient because it provides an explicit representation for the closure of the parameter space under mean-value parametrization. Specifically, consider the set

$$\mathrm{V}_{\mathrm{A},\geq 0} = \left\{ \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \colon \mathbf{x}^{\mathbf{z}^+} = \mathbf{x}^{\mathbf{z}^-}, \forall \mathbf{z} \in \mathrm{kernel}(\mathrm{A}) \cap \mathbb{Z}^{\mathcal{I}} \right\}$$

consisting of the real-valued non-negative solutions of a system of polynomial equations specified by the lattice points in the orthogonal complement of $\mathcal{M}$. The set $\mathrm{V}_{\mathrm{A},\geq 0}$ is called a *toric variety*

(Sturmfels, 1996; Diaconis and Sturmfels, 1998) and defines a closed, smooth hyper-surface of point in the non-negative orthant of $\mathbb{R}^{\mathcal{I}}$. The most notable example of toric varieties arising from hierarchical log-linear models is the surface of independence for a 2×2 and the model of independence (Fienberg and Gilbert, 1970). The set $\mathrm{V}_{\mathrm{A},\geq 0}$ offers a more advantageous description of the parameter space under mean value parametrization than the natural parametrization or the u-term representation, which parametrize only log-linear models for which the cell mean vector has strictly positive coordinates. In contrasts, points in $\mathrm{V}_{\mathrm{A},\geq 0}$ with some zero coordinates correspond to boundary log-linear models, i.e. statistical models described by probability distributions expressible in exponential form whose cell mean vectors belong to the boundary of $\overline{M}$. To complete the geometric characterization of log-linear models and extended maximum likelihood estimation, we introduce another geometric object that plays a fundamental role in the geometry of log-linear model, namely the polytope

$$\mathrm{P}_{\mathbf{t}} = \left\{ \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \colon \mathbf{t} = \mathrm{A}\mathbf{x} \right\},$$

consisting of the set of all possible cell mean values whose margin match the observed sufficient statistics $\mathbf{t}$. The set of lattice points inside $\mathrm{P}_{\mathbf{t}}$, called the *fiber* at $\mathbf{t}$, corresponds to the support of the conditional distribution of the the tables given the value $\mathbf{t}$ for the sufficient statistics, often known as the "exact distribution".

Since both the MLE and extended MLE satisfy the moment equations, $\widehat{\mathbf{m}}^{\mathrm{e}}$ is a point of $\mathrm{P}_{\mathbf{t}}$ (in fact, it is a point in the relative interior of $\mathrm{P}_{\mathbf{t}}$). The marginal cone C_{A}, the variety $\mathrm{V}_{\mathrm{A},\geq 0}$ and the polytope $\mathrm{P}_{\mathbf{t}}$ are the three main geometric object that fully describe the geometry of log-linear models and their closure. Their relationships are summarized in the following theorem. See Rinaldo (2006) and Geiger et al. (2006) for the proofs of these statements.

Theorem 2.6.

i. $\overline{M} = \mathrm{V}_{\mathrm{A},\geq 0}$*;*

ii. $\mathrm{V}_{\mathrm{A},\geq 0}$ *and* C_{A} *are homeomorphic and, for any* $\mathbf{m} \in \mathrm{V}_{\mathrm{A},\geq 0}$, $\mathrm{supp}(\mathbf{m})$ *is a facial set of* C_{A} *;*

iii. for any $\mathbf{t} = \mathrm{A}\mathbf{n}$, $\{\widehat{\mathbf{m}}^{\mathrm{e}}\} = \mathrm{V}_{\mathrm{A},\geq 0} \cap \mathrm{P}_{\mathbf{t}}$ *and* $\widehat{\mathbf{m}}^{\mathrm{e}} \in \mathrm{ri}(\mathrm{P}_{\mathbf{t}})$.

The first results says that the toric variety consists exactly of the set of all cell mean vectors for the log-linear model specified by the row range of A and its point-wise limit closure (i.e. the "boundary" models). The second results says that there exists a one-to-one correspondence between the expected values of the sufficient statistics and the cell mean vectors. This implies that, for any observable value of the margins $\mathbf{t} = \mathrm{A}\mathbf{n}$, there exists one and only one point $\widehat{\mathbf{m}}$ in $\mathrm{V}_{\mathrm{A},\geq 0}$ such that $\mathrm{A}\mathbf{n} = \mathrm{A}\widehat{\mathbf{m}}$. This point is the MLE if $\mathbf{t}$ is in the interior of C_{A}, in which case $\mathrm{supp}(\widehat{\mathbf{m}}) = \mathcal{I}$ or the extended MLE if $\mathbf{t}$ is on the boundary of the marginal cone, in which case $\mathrm{supp}(\widehat{\mathbf{m}}) = \mathcal{F}$, with $\mathcal{F}$ being the facial set determined by $\mathbf{t}$. The conditions on the support of the points in $\mathrm{V}_{\mathrm{A},\geq 0}$ implies that the additional distributions parametrized, in a mean value sense, by the boundary of $\mathrm{V}_{\mathrm{A},\geq 0}$ (and hence of C_{A}) are supported on facial sets of the marginal cone. The last statement is a geometrical representation of the result that the (extended) MLE is the only point in $\overline{M}$ that satisfies the moment equations.

Overall, the results described in this section allow to identify boundary log-linear models in many fashions: analytically by maximizing the log-likelihood function, geometrically by describing points on the boundary of both $\mathrm{V}_{\mathrm{A},\geq 0}$ and C_{A} and combinatorially through the facial sets of the

marginal cone. Not only do they provide useful theoretical characterizations of extended exponential families and extended MLE but also they give a full description of all possible patterns of sampling zeros leading to a nonexistent MLE and the proposal of alternative inferential procedure to deal with tables with a nonexistent MLE (see Section 6).

In fact, the condition that the MLE exists if and only if $\mathbf{t} \in \text{relint}(\text{C}_\text{A})$ can be translated into the equivalent statement that the MLE does not exists if and only if $\text{supp}(\mathbf{n}) \subseteq \mathcal{F}$ for some facial set $\mathcal{F}$ of C_A. In other words, the "forbidden" configurations of sampling zeros causing the MLE to be undefined are precisely the complements of the facial set. We call these configurations of zero counts *likelihood zeros*. Examples of likelihood zeros for different log-linear models are given in Rinaldo (2006) and Eriksson et al. (2006). The number of likelihood zero configurations is directly related to the combinatorial complexity of the marginal cone which, for most log-linear models, appears to grow super-exponentially with the number of variable K and the number of categories I_k, $k = 1, \ldots, K$, (see the computational study in Eriksson et al., 2006). In Section 4 we will address this problem and derive efficient methods for computing the likelihood zeros, the extended MLE and the dimension of the statistical model identified by it.

3 Log-Linear Models Subspaces

Although log-linear models are defined by generic linear manifolds of $\mathbb{R}^{\mathcal{I}}$, in practice it is customary to consider only certain, rather constrained, classes of linear subspaces. These subspaces, which are also characteristic of ANOVA models and experimental design, present considerable advantages in terms of interpretability and ease of computation. In this section we describe various, equivalent characterizations of the *factor-interaction subspaces* and the *subspaces of interactions* described in Darroch and Speed (1983) using combinatorics, linear algebra and group theory. We will also present algorithms for optimal sparse representations of the relevant design matrices that will prove useful in the next section for efficient computation of the extended MLE and of the number of estimable parameters.

Let $\mathcal{K} = \{1, 2, \ldots, K\}$ the set of labels of the variables of interest and let $2^{\mathcal{K}}$ be the power set of $\mathcal{K}$. There is a natural partial order in $2^{\mathcal{K}}$ induced by the operation of taking subset inclusion which makes $2^{\mathcal{K}}$ into a boolean lattice. A hypergraph $\mathcal{H}$ on the finite base set $\mathcal{K}$ is a subset of $2^{\mathcal{K}}$ consisting of a class of subsets of $\mathcal{K}$ called *hyperedges*. An abstract simplicial complex Δ on $\mathcal{K}$ is a hypergraph on $\mathcal{K}$ such that $h \subset d$ for some $d \in \Delta$ implies $h \in \Delta$. An antichain is a subset $\mathcal{A}$ of $2^{\mathcal{K}}$ such that, for any $a_i, a_j \in \mathcal{A}$, $a_i \not\subseteq a_j$. Since K is finite, there is a one-to-one correspondence between antichains and simplicial complexes of $2^{\mathcal{K}}$. Namely, for any simplicial complex Δ, the set of all maximal hyperedges of Δ forms an antichain. Conversely, given any antichain $\mathcal{A}$, the class of sets

$$\Delta = \{d \subset \mathcal{K} \colon d \subseteq a, \text{ for some } a \in \mathcal{A}\}$$

forms a simplicial complex. Because of this correspondence, any simplicial complex Δ will be identified with the associated antichain of maximal hyperedges, called the *facets* of Δ, while the non-maximal hyperedges will be called its *faces*.

The definition of log-linear models given below is essentially based on the combinatorial structure of the power set $2^{\mathcal{K}}$ and provides a formal justification of the traditional notation (see, for example, Bishop et al., 1975, Chapter 3) of identifying log-linear models, and in particular hierarchical log-linear models, with classes of subsets of $\mathcal{K}$, often called *generating classes*. Such

characterizations are virtually identical to analysis of variance models, a connection that has been known for a long time (see for example Haberman, 1974; Bishop et al., 1975), so that the log-linear subspaces derived here are the same orthogonal decompositions of $\mathbb{R}^{\mathcal{I}}$ defined by Darroch and Speed (1983). In fact, the contents of Section 3.1 of the present chapter can be considered as an alternative derivation of some results of Darroch and Speed (1983).

Definition 3.1. A log-linear model is a hypergraph $\mathcal{H}$ on $\mathcal{K}$ and a hierarchical log-linear model is a simplicial complex Δ on $\mathcal{K}$.

Graphical models form a subclass of hierarchical log-linear models (Whittaker, 1990; Lauritzen, 1996). To each hypergraph $\mathcal{H}$ on $\mathcal{K}$, it is always possible to associate its *2-section* $\mathcal{G}(\mathcal{H})$ (see, e.g., Berge, 1989), which is a graph with vertex set $V = \{1, \ldots, K\}$ and edge set consisting of all unordered pairs $(i, j) \subset V$ such that $(i, j) \subseteq h$ for some $h \in \mathcal{H}$. In the statistical literature, the 2-section graph is also called *interaction graph*. A hypergraph $\mathcal{H}$ is called *graphical* if its hyperedges are the cliques of its 2-section $\mathcal{G}(\mathcal{H})$.

Definition 3.2. A hierarchical log-linear model Δ is graphical if the facets of Δ form a graphical hypergraph.

The log-linear subspace $\mathcal{M}$ associated to a hierarchical log-linear model Δ will be indicated as $\mathcal{M}_\Delta$ and, similarly, $\mathcal{M}_\mathcal{H}$ will denote the log-linear subspace corresponding to a generic log-linear model specified by the hypergraph $\mathcal{H}$. Unless otherwise specified, we will always assume that, for a generic log-linear model $\mathcal{H}$, $\bigcup_{h\in\mathcal{H}} = \mathcal{K}$, so that all the K factors are always included. Each hyperedge of $\mathcal{H}$ will be given straightforward interpretation, using the language of analysis of variance. Specifically, the hyperedges h are called $|h|$-factor interaction term or interaction term of order $|h| - 1$. If $|h| = 1$, then h is a main effect. If $|h| = 0$, then h is the grand mean.

In this document, for a hierarchical log-linear model Δ on K factors, the generating classes will be represented by a list of factors enclosed in squared brackets, i.e. $[1]$ indicates the main effect for the factor labeled as "1" and $[12]$ the interaction between factor "1" and "2". This notation is consistent with the one used in the statistical literature (e.g. Bishop et al., 1975; Fienberg, 1980).

Example 3.3 (Hierarchical log-linear models). $\Delta = [1][2][3]$ is the model of mutual independence of the three factors and $\Delta = [12][23]$ denotes the model of conditional independence of factor "1" and "3" given factor "2" (a decomposable model; see Section 5.2). The simplest non-graphical model is the model of no-3-factor effect $\Delta = [12][23][13]$, which will be used as a test set for many examples here. In fact, for a K-way table, the largest hierarchical log-linear model is the model of no-K-factor effect, represented by the simplicial complex on K nodes whose $K - 1$ facets form the set of all possible distinct subsets of $\mathcal{K}$ with cardinality $K - 1$. The simplest example of a graphical non-decomposable (and non-reducible) model is the 4-cycle model on 4 factors, $\Delta = [12][23][34][14]$. ■

For a log-linear model $\mathcal{H}$, we will represent the log-linear subspace $\mathcal{M}_\mathcal{H}$ as the direct sum of orthogonal subspaces of $\mathbb{R}^{\mathcal{I}}$, each determined by a subset of $\mathcal{K}$,

$$\mathcal{M}_\mathcal{H} = \bigoplus_{h\in\mathcal{H}} \mathcal{M}_h, \tag{10}$$

where $\mathcal{M}_h \perp \mathcal{M}_{h'}$ for $h, h' \subseteq \mathcal{K}$ with $h \neq h'$. For a hierarchical log-linear model Δ, the above expression specializes to

$$\mathcal{M}_\Delta = \bigoplus_{\{h\subseteq d:\ d\in\Delta\}} \mathcal{M}_h. \tag{11}$$

In the remainder of the section various equivalent constructions of the subspaces involved in Equations (10) and (11) are provided. These constructions will lead to the same class of subspaces described in Darroch and Speed (1983).

3.1 Combinatorial Derivation

In this section we derive log-linear model subspaces is obtained by establishing multiple correspondences among the combinatorial structure of the subsets of $\mathcal{K}$, the properties of certain classes of partitions of $\mathcal{I}$ and the decomposition of $\mathbb{R}^{\mathcal{I}}$ into direct sums of orthogonal linear subspaces. The techniques and most of the results utilized here are adapted from Bailey (2004).

Let $h \subset \mathcal{K}$ and define the equivalence relation $\overset{h}{\sim}$ on $\mathcal{I}$ given by

$$i \overset{h}{\sim} j \iff i_h = j_h,$$

for all $i, j \in \mathcal{I}$. The equivalence classes of $\overset{h}{\sim}$ in turn define a partition $p(h)$ of $\mathcal{I}$ into $d_h = \prod_{k \in h} I_k$ subsets of equal cardinality $n_h = \prod_{k \notin h} I_k = \frac{I}{d_h}$. Explicitly, the equivalence classes are the sets $\{\mathcal{H}_j \colon j \in \mathcal{I}_h\}$, where $\mathcal{H}_j = \{i \in \mathcal{I} \colon i_h = j\}$ and $\mathcal{I}_h = \bigotimes_{k \in h} \mathcal{I}_k$, so that $\mathcal{I} = \uplus_{j \in \mathcal{I}_h} \mathcal{H}_j$. It should be noted that an identical partitioning argument is utilized for computing an orthogonal basis of the sampling subspace for the product-multinomial sampling scheme in Section 2. In fact, d_h is the number of entries of the h-margins of any function $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$ and n_h is the number of cells to be summed over in order to compute each such entry.

Partitions whose classes have constant size are said to be *uniform*. All the partitions defined in the way described above are, by construction, uniform. For any such partition $p(h)$, the relation matrix $\mathrm{R}_h \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ is the incidence matrix of the equivalence classes of $\overset{h}{\sim}$ over $\mathcal{I}$ and is defined as

$$\mathrm{R}_h(i, j) = \begin{cases} 1 & \text{if } \; i \overset{h}{\sim} j \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The relation matrix R_h has the explicit form, of easy verification,

$$\mathrm{R}_h = \bigotimes_{k=1}^{K} \begin{cases} \mathrm{I}_k & \text{if } \; k \in h \\ \mathrm{J}_k & \text{if } \; k \notin h, \end{cases}$$

where I_k is the I_k-dimensional identity matrix and J_k the I_k-dimensional matrix containing all 1's.

Associated to each $p(h)$ is the subspace $\mathcal{W}_h \subset \mathbb{R}^{\mathcal{I}}$ consisting of all functions on the multi-index set $\mathcal{I}$ that depend on $i \in \mathcal{I}$ only through i_h, i.e.

$$\mathcal{W}_h = \left\{ f \in \mathbb{R}^{\mathcal{I}} \colon f(i) = f(j) \quad \text{if} \quad i \overset{h}{\sim} j \right\}, \tag{13}$$

where $\mathcal{W}_\emptyset = \mathbf{1}$ and $\mathcal{W}_{\mathcal{K}} = \mathbb{R}^{\mathcal{I}}$. In Darroch and Speed (1983) such spaces are called *factor-interaction subspaces*. Let P_h be the orthogonal projection onto $\mathcal{W}_h$. Then, for any $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, $\mathbf{y} = \mathrm{P}_h \mathbf{x}$ is a vector whose i_h-margins are constant and equal to the average of the corresponding coordinates of $\mathbf{x}$, i.e.

$$\mathbf{y}(i) = \frac{1}{n_h} \sum_{\{j \in \mathcal{I} \colon i \overset{h}{\sim} j\}} \mathbf{x}(j).$$

As a consequence of the uniformity of the classes of $p(h)$, the projection matrix P_h has a very simple form, obtained as

$$\mathrm{P}_h = \frac{1}{n_h}\mathrm{R}_h = \bigotimes_{k=1}^{K} \begin{cases} \mathrm{I}_k & \text{if } \ k \in h \\ \frac{1}{I_k}\mathrm{J}_k & \text{otherwise.} \end{cases} \tag{14}$$

Let $P^{\mathcal{K}} = \{p(h) \colon h \in 2^{\mathcal{K}}\}$, where $p(\emptyset)$ consists of the entire set $\mathcal{I}$, and $p(\mathcal{K})$ has as many classes as cells. For different partitions $p(h)$ and $p(h')$, write $p(h) \preccurlyeq p(h')$ if every $p(h)$-class is a subset of some $p(h')$-class. The relation $\preccurlyeq$ satisfies the properties of reflexivity, antisymmetry and transitivity, so that $P^{\mathcal{K}}$ is a poset with respect to the relation $\preccurlyeq$. Similarly, let $\mathcal{W}^{\mathcal{K}} = \{\mathcal{W}_h \colon h \in 2^{\mathcal{K}}\}$, where the subspaces are defined as in (13). By construction, $\mathcal{W}^{\mathcal{K}}$ is partially ordered by inclusion, hence it is a poset with respect to the inclusion operator.

Lemma 3.4. *The three posets $\mathcal{W}^{\mathcal{K}}$, $2^{\mathcal{K}}$ and $P^{\mathcal{K}}$ are isomorphic lattices:*

$$\mathcal{W}_{h'} \subseteq \mathcal{W}_h \iff h' \ \subseteq h \iff p(h') \succcurlyeq p(h) \tag{15}$$

for all $h, h' \in 2^{\mathcal{K}}$.

Proof. The equivalences follow by construction. Using (15) and the fact that $2^{\mathcal{K}}$ is a lattice or direct verification, it follows that both $P^{\mathcal{K}}$ and $\mathcal{W}^{\mathcal{K}}$ are lattices. ■

As indicated in the equivalence (15), the relation $\preccurlyeq$ on $P^{\mathcal{K}}$ the is the "opposite" of the inclusion relation $\subseteq$ on both $2^{\mathcal{K}}$ and $\mathcal{W}^{\mathcal{K}}$; in fact, subsets of h and linear subspaces of $\mathcal{W}_h$ in $\mathcal{W}^{\mathcal{K}}$ induce coarser partitions than $p(h)$. The $\hat{0}$ and $\hat{1}$ elements of $P^{\mathcal{K}}$ are $p(\emptyset)$ and $p(\mathcal{K})$, respectively, while $\hat{0}$ and $\hat{1}$ elements of $\mathcal{W}^{\mathcal{K}}$ are $\mathbf{1}_{\mathcal{I}}$ and $\mathbb{R}^{\mathcal{I}}$, respectively. Furthermore, the relations of the next Corollary can be verified for the least upper bound and biggest lower bound on $P^{\mathcal{K}}$.

Corollary 3.5. *For any $h_1, h_2, h_3 \in 2^{\mathcal{K}}$:*

$$\begin{aligned} \mathcal{W}_{h_3} = \mathcal{W}_{h_1} + \mathcal{W}_{h_2} &\iff h_3 = h_1 \cup h_2 \iff p(h_3) = p(h_1) \wedge p(h_1) \\ \mathcal{W}_{h_3} = \mathcal{W}_{h_1} \cap \mathcal{W}_{h_2} &\iff h_3 = h_1 \cap h_2 \iff p(h_3) = p(h_1) \vee p(h_1) \end{aligned} \tag{16}$$

Because of the isomorphism of Lemma 3.4, combinatorial properties of $P^{\mathcal{K}}$ and $2^{\mathcal{K}}$ translates into geometric properties of the subspaces in $\mathcal{W}^{\mathcal{K}}$. In order to illustrate these properties we will use the correspondence between orthogonal partitions and geometric orthogonality. Specifically, two partitions $p(h)$ and $p(h')$ are said to be orthogonal if $\mathrm{P}_h\mathrm{P}_{h'} = \mathrm{P}_{h'}\mathrm{P}_h$. Linear subspaces whose projection matrices commute, like the ones associated with orthogonal partitions, are called *geometrically orthogonal*. As it turns out, orthogonality of all the partitions making up $P^{\mathcal{K}}$ is precisely the required feature to obtain a combinatorial decomposition of $\mathbb{R}^{\mathcal{I}}$ into orthogonal subspaces indexed by elements of $2^{\mathcal{K}}$, as in Equations (10) and (11).

Lemma 3.6. *For any $h, h' \in 2^{\mathcal{K}}$:*

1. *$p(h)$ and $p(h')$ are orthogonal, hence $\mathcal{W}_h$ and $\mathcal{W}_{h'}$ are geometrically orthogonal;*
2. *$\mathcal{W}_h \cap \left(\mathcal{W}_{p(h)\vee p(h')}\right)^{\perp}$ is orthogonal to $\mathcal{W}_{h'}$.*

Proof. Lemma 6.4 and Lemma 9.1 in Bailey (2004). ■

The geometric orthogonality of the vector spaces $\mathcal{W}_h$, $h \in 2^{\mathcal{K}}$, allows for a recursive decomposition of every $\mathcal{W}_h$ into orthogonal subspaces computed by intersecting $\mathcal{W}_h$ itself with the orthogonal complement of all the smaller subspaces it contains.

Theorem 3.7. *For any $h \in 2^{\mathcal{K}}$, define:*

$$\mathcal{U}_h = \mathcal{W}_h \cap \left(\sum_{\{h' \in 2^{\mathcal{K}} \colon p(h') \succ p(h)\}} \mathcal{W}_{h'} \right)^{\perp}. \tag{17}$$

Then:

1. *for any $h, h' \in 2^{\mathcal{K}}$, with $h \neq h'$ the subspaces $\mathcal{U}_h$ and $\mathcal{U}_{h'}$ are orthogonal to each other;*
2. *for each $h \in 2^{\mathcal{K}}$*

$$\mathcal{W}_h = \bigoplus_{\{h' \in 2^{\mathcal{K}} \colon p(h') \succcurlyeq p(h)\}} \mathcal{U}_{h'}. \tag{18}$$

Proof. Theorem 6.7 in Bailey (2004). ■

In virtue of (15), $\{h' \in 2^{\mathcal{K}} \colon p(h') \succcurlyeq p(h)\} = \{h' \in 2^{\mathcal{K}} \colon h' \subseteq h\}$, so that equation (18), along with $\mathcal{W}_{\mathcal{K}} = \mathbb{R}^{\mathcal{I}}$, gives:

Corollary 3.8. *For any $h \in 2^{\mathcal{K}}$:*

$$\mathcal{W}_h = \bigoplus_{h' \subseteq h} \mathcal{U}_{h'}. \tag{19}$$

In particular:

$$\mathbb{R}^{\mathcal{I}} = \bigoplus_{h' \subseteq \mathcal{K}} \mathcal{U}_{h'}. \tag{20}$$

The orthogonal subspaces $\mathcal{U}_h$ decomposing the $\mathcal{W}_h$'s are the *subspaces of interactions* in Darroch and Speed (1983) (see also Lauritzen, 1996, Appendix B). Bailey (2004) calls them *strata* instead.

In the rest of the section, the combinatorial structure of $P^{\mathcal{K}}$ will be exploited We will take advantage of the combinatorial structure of $P^{\mathcal{K}}$ to derive formulas for the projection matrices onto the factor subspaces and, consequently, a fully description of the subspaces of interactions.

For any two partitions $p(h)$ and $p(h')$ in the lattice $P^{\mathcal{K}}$, define in $\mathbb{R}^{P_{\mathcal{K}} \times P_{\mathcal{K}}}$ the *zeta function*

$$\zeta\left(p(h), p(h')\right) = \begin{cases} 1 & \text{if } \; p(h) \preccurlyeq p(h') \\ 0 & \text{otherwise.} \end{cases}$$

Since $P^{\mathcal{K}}$ is finite, it is possible to assign an ordering to its elements in such a way that $p(h)$ comes before $p(h')$ if $p(h) \preccurlyeq p(h')$. The zeta function belongs to the incidence algebra (see, for example, Stanley, 1997) of $P^{\mathcal{K}}$, which is isomorphic to the algebra of upper triangular matrices. Provided such an ordering of the elements of $P^{\mathcal{K}}$ has been fixed once and for all, the zeta function can be represented as an upper triangular matrix with diagonal entries all equal to 1. The inverse of such a matrix is therefore well defined and is isomorphic to another function in the corresponding incidence algebra, called the *Möbius function* $\boldsymbol{\mu}$ of the poset $\left(P^{\mathcal{K}}, \preccurlyeq\right)$. Using the Möbius function it is possible to obtain a combinatorial representation of both the projector and the dimension of the subspaces of interactions $\mathcal{U}_h$, $h \in 2^{\mathcal{K}}$.

Theorem 3.9. *For $h \in 2^{\mathcal{K}}$, the projector S_h onto the subspace $\mathcal{U}_h$ defined in (17) is given by*

$$\mathrm{S}_h = \sum_{h' \in 2^{\mathcal{K}}} \boldsymbol{\mu}\left(p(h), p(h')\right) \mathrm{P}_{h'} \tag{21}$$

and the dimension of S_h is

$$\dim(\mathcal{U}_h) = \sum_{h' \in 2^{\mathcal{K}}} \boldsymbol{\mu}\left(p(h), p(h')\right) d_{h'}.$$

Proof. This result follows from Theorem 6.9 in Bailey (2004) after noting that each $p(h)$ form an *orthogonal block structure* (Bailey, 2004, Chapter 6). ■

Using the last theorem it is possible to derive a formula for S_h.

Corollary 3.10.

$$\mathrm{S}_h = \bigotimes_{k \in \mathcal{K}} \mathrm{S}_k^h \tag{22}$$

where

$$\mathrm{S}_k^h = \begin{cases} \mathrm{I}_k - \frac{1}{I_j}\mathrm{J}_k & \text{if} \quad k \in h \\ \frac{1}{Ij}\mathrm{J}_k & \text{otherwise.} \end{cases}$$

Note that the last equation match the formulas for the log-linear models of a 3-way table found in Knuiman and Speed (1988).

Proof. Because there is an order-reversing correspondence between the lattices $\left(2^{\mathcal{K}}, \subseteq\right)$ and $\left(P^{\mathcal{K}}, \preccurlyeq\right)$, as indicated in (15), $\boldsymbol{\mu}\left(p(h), p(h')\right) = \boldsymbol{\mu}(h', h)$, where it is clear that the first Möbius function refers to $P^{\mathcal{K}}$ and the second to $2^{\mathcal{K}}$. Next, the Möbius function for the lattice $2^{\mathcal{K}}$ is (see Stanley, 1997, page 118)

$$\boldsymbol{\mu}(h', h) = \begin{cases} (-1)^{|h \setminus h'|} & \text{if} \quad h' \subseteq h \\ 0 & \text{otherwise.} \end{cases}$$

Hence, equation (21) becomes

$$\mathrm{S}_h = \sum_{\{h' \in 2^{\mathcal{K}} \colon h' \subseteq h\}} (-1)^{|h \setminus h'|} \mathrm{P}_{h'}. \tag{23}$$

It is worth noting that the this is identical to Equation (4.3) in Darroch and Speed (1983) and Equation (B.15) in Lauritzen (1996).

Next, assume for simplicity and without loss of generality that $h = \{1, \ldots, |h|\}$ (in fact, the whole construction is invariant under permutations of the K factors). Because of the linearity of the tensor product an induction argument can be used to show that

$$\begin{aligned} \textstyle\bigotimes_{k=1}^{|h|} \mathrm{I}_k - \frac{1}{I_k}\mathrm{J}_k &= \textstyle\sum_{\boldsymbol{\delta} \in \{0,1\}^{|h|}} (-1)^{\sum_k \delta_k} \bigotimes_{k=1}^{|h|} \left(\delta_k \mathrm{I}_k + (1 - \delta_k)\frac{1}{I_k}\mathrm{J}_k\right) \\ &= \textstyle\sum_{h' \subseteq h} (-1)^{|h'|} \bigotimes_{k=1}^{|h|} \left(\delta_k \mathrm{I}_k + (1 - \delta_k)\frac{1}{I_k}\mathrm{J}_k\right). \end{aligned}$$

Then, using the previous equality and the tensor product form of the projection matrices P_h given in (14), a more explicit representation of (23) can be obtained as

$$\begin{aligned}
\mathrm{S}_h &= \textstyle\sum_{\{h' \subseteq h\}} (-1)^{|h \setminus h'|} \mathrm{P}_{h'} \\
&= \textstyle\sum_{\{h' \subseteq h\}} (-1)^{|h'|} \mathrm{P}_{h \setminus h'} \\
&= \left[\textstyle\sum_{h' \subseteq h} (-1)^{|h'|} \bigotimes_{k=1}^{|h|} \left(\delta_k \mathrm{I}_k + (1 - \delta_k) \frac{1}{I_k} \mathrm{J}_k \right) \right] \bigotimes_{k > |h|} \frac{1}{I_k} \mathrm{J}_k \\
&= \bigotimes_{k \in h} \left(\mathrm{I}_k - \frac{1}{I_k} \mathrm{J}_k \right) \bigotimes_{k > |h|} \frac{1}{I_k} \mathrm{J}_k,
\end{aligned}$$

which gives, possibly accounting for permutations of the factors, Equation (22). ■

Formulas for the dimension of the subspaces $\mathcal{U}_h$, which again utilizes the Möbius function, will be derived later in Corollary 3.15. By setting $\mathcal{M}_h = \mathcal{U}_h$ for each $h \subseteq \mathcal{K}$, the desired representation of Equations (10) and (11) is obtained.

Definition 3.11. The log-linear subspace associated to a log-linear model $\mathcal{H}$ is defined to be

$$\mathcal{M}_{\mathcal{H}} = \bigoplus_{h \in \mathcal{H}} \mathcal{U}_h. \tag{24}$$

As for hierarchical log-linear models, Equation (19) allows for the following refinement of the previous definition:

Definition 3.12. The log-linear subspace associated to a hierarchical log-linear model with generating classes Δ is

$$\mathcal{M}_{\Delta} = \sum_{d \in \Delta} \mathcal{W}_d = \bigoplus_{\{h \subseteq d \colon d \in \Delta\}} \mathcal{U}_h. \tag{25}$$

3.2 Matrix Algebra Derivation

This section provides efficient algorithms for building design matrices for the factor-interaction subspaces $\mathcal{W}_h$ and the subspaces of interactions $\mathcal{U}_h$, $h \in 2^{\mathcal{K}}$. The columns of such matrices will span vector subspaces satisfying the defining Equations (24) and (25) of log-linear and hierarchical log-linear models, respectively. A significant portion of the material is derived from Fienberg et al. (1980).

3.2.1 Bases for $\mathcal{U}_h$: Contrast Bases

Given a log-linear model $\mathcal{H}$, bases for the subspaces $\mathcal{U}_h = \mathcal{M}_h$, with $h \in \mathcal{H}$ will be defined and computed. The term *contrast bases* is appropriate because they indeed correspond to contrasts in models of analysis of variance. Using Birch's notation (see, in particular, Bishop et al., 1975), the design matrix for $\mathcal{U}_h$ will encode to the *u-terms* corresponding to the $|h|$-order interactions among the factors in h.

For each term $h \subseteq \mathcal{K}$ and factor $k \in \mathcal{K}$, define the matrix

$$\mathrm{U}_k^h = \begin{cases} \mathrm{Z}_k & \text{if} \quad k \in h \\ \mathbf{1}_k & \text{if} \quad k \notin h, \end{cases}$$

where Z_k is a $I_k \times (I_k - 1)$ matrix with entries

$$\mathrm{Z}_k = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & -1 \end{pmatrix}, \tag{26}$$

and $\mathbf{1}_k$ is the I_k-dimensional column vector of 1's. Let

$$\mathrm{U}_h = \bigotimes_{k=1}^{K} \mathrm{U}_k^h. \tag{27}$$

Since the elements of U_k^h are -1,0 and 1, U_h has entries that can only be -1,0 and 1. Additional properties of the design matrices U_h and hence of the subspaces spanned by their columns are given in the next Lemma.

Lemma 3.13.

i. For every $h, h' \in 2^{\mathcal{K}}$, with $h \neq h'$, the columns of U_h are linearly independent and $\mathrm{U}_h^\top \mathrm{U}_{h'} = 0$;

ii. $\mathbb{R}^{\mathcal{I}} = \bigoplus_{h \in 2^{\mathcal{K}}} \mathcal{R}(\mathrm{U}_h)$;

iii. for any $h \in 2^{\mathcal{K}}$, $\mathcal{R}(\mathrm{U}_h) = \mathcal{U}_h$, where $\mathcal{U}_h$ is the subspace of interactions for the factors in h.

Proof. Part *i.*: the first statement follows from the fact that the columns of Z_k^h are independent for each k and h and U_h has dimension $\left(\prod_{k=1}^K I_k\right)$. As for the second statement, without loss of generality, we can assume that there exists a factor k such that $k \in h$ and $k \notin h'$. Then, $\mathrm{Z}_k^h = \mathrm{C}_k$ and $\mathrm{Z}_k^{h'} = \mathbf{1}_k$, so $(\mathrm{Z}_k^h)^\top \mathrm{Z}_k^{h'} = 0$, hence the result.
Part *ii.*: by *i.*, the subspaces $\mathcal{R}(\mathrm{U}_h)$, are orthogonal (hence the direct sum notation is well defined) and $\dim \mathcal{R}(\mathrm{U}_h) = \prod_{j \in h}(I_j - 1)$. Therefore:

$$\dim\left(\bigoplus_{h \in 2^{\mathcal{K}}} \mathcal{R}(\mathrm{U}_h)\right) = \sum_{h \in 2^{\mathcal{K}}} \prod_{j \in h} (I_j - 1) = \prod_{k \in \mathcal{K}} I_k,$$

where the last equality follows from Lemma 3.18 and the fact that, for $h = \emptyset$, $\dim(\mathrm{U}_h) = 1$ since $\mathrm{U}_\emptyset = \mathbf{1}_{\mathcal{I}}$.
Part *iii.*: it suffices to show $\mathrm{S}_h \mathrm{U}_h = \mathrm{U}_h$ and $\mathrm{S}_h \mathrm{U}_{h'} = 0$ for $h \neq h'$, where S_h is the projection matrix onto $\mathcal{U}_h$ as in Equation (22). This implies $\mathcal{R}(\mathrm{U}_h) \subseteq \mathcal{U}_h$ and the results follow from the fact that the inclusion cannot be strict because of the orthogonal decompositions of $\mathbb{R}^{\mathcal{I}}$ as in *ii.* and Equation (20). It is easy to see that $\mathrm{S}_k^h \mathrm{U}_k^h = \mathrm{U}_k^h$ and, for any $h \neq h'$ with $k' \in h' \backslash h$, $\mathrm{S}_{k'}^{h'} \mathrm{U}_{k'}^h = \mathrm{S}_{k'}^h \mathrm{U}_{k'}^{h'} = 0$. Therefore,

$$\mathrm{S}_h \mathrm{U}_h = \bigotimes_{k=1}^{K} \mathrm{S}_k^h \mathrm{U}_k^h = \mathrm{U}_h \qquad \text{and} \qquad \mathrm{S}_h \mathrm{U}_{h'} = \bigotimes_{k=1}^{K} \mathrm{S}_k^h \mathrm{U}_k^{h'} = 0$$

for any $h \neq h'$. ■

Note that part *ii.* yields a decomposition of $\mathbb{R}^{\mathcal{I}}$ analogous to Equation (20).

An alternative way of computing the contrast basis, which might be advantageous from the computational viewpoint, is to replace the matrix Z_k in (26) by the equivalent matrix, having the same dimension,

$$\mathrm{C}_k = \begin{pmatrix} 1 & 1 & \dots & 1 \\ -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & -1 \end{pmatrix}. \tag{28}$$

It is immediate to see that $\mathcal{R}(\mathrm{Z}_k) = \mathcal{R}(\mathrm{C}_k)$ and hence part *.i* and *ii.* of Lemma 3.13 still hold.

It is possible to generate the contrast bases design matrices in computationally efficient ways, which is especially useful when dealing with very high-dimensional problems. Section 3.4.2 in the Appendix gives algorithms for generating the matrices U_h one row at a time and also for storing them with minimal memory allocation requirements.

3.2.2 Bases for $\mathcal{W}_h$: Marginal Bases

Although bases for hierarchical models can be computed using contrast bases, there is a different construction for the hierarchical log-linear subspaces which produces sparse, redundant bases called *marginal bases*. The use of marginal bases is very common, mainly because of ease of interpretability: such bases in fact induce minimal sufficient statistics which coincides the marginal table sums. In addition, as it is shown in Section 5.2, design matrices defined with marginal basis allow to identify quite straightforwardly some cases in which the MLE is undefined.

For any $h \in 2^{\mathcal{K}}$ let

$$\mathrm{W}_k^h = \begin{cases} \mathrm{I}_{I_k} & \text{if} \quad k \in h \\ \mathbf{1}_k & \text{if} \quad k \notin h, \end{cases}$$

and

$$\mathrm{W}_h = \bigotimes_{k=1}^{K} \mathrm{W}_k^h, \tag{29}$$

where I_{I_k} denotes the I_k-dimensional identity matrix.

Let $\mathcal{I}_h = \prod_{k \in h} \mathcal{I}_k$, so that $\mathcal{I}_h$ contains the indexes for the columns of W_h, where the usual lexicographic ordering introduced in Section 1.1 is assumed. The 0-1 $\left(\prod_{k=1}^{K} I_k\right) \times \left(\prod_{k \in h} I_k\right)$-dimensional matrix W_h enjoys the following properties, which can be easily verified:

1. it is full-column-rank: $\mathrm{rank}(\mathrm{W}_h) = \left(\prod_{k \in h} I_k\right)$;
2. for each $j \in \mathcal{I}_h$, the column indexed by j have zero entries except in the coordinates $i \in \mathcal{I}$ such that $i_h = j$.
3. $\mathrm{W}_h^\top \mathrm{W}_h = 0$;

4. $\mathbf{1}_{\mathcal{I}}\mathrm{W}_h^\top = \mathbf{1}_h n_h$, where $\mathbf{1}_h$ is the d_h-dimensional vector containing only 1's, with n_h and d_h defined at the beginning of Section 3.1.

It is straightforward to see that, by construction, $\mathcal{R}(\mathrm{W}_h) = \mathcal{W}_h$, with $\mathcal{W}_h$ being the factor-interaction subspace defined as in (13). Notice also, that, for $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, the linear transformation $\mathrm{W}_h^\top \mathbf{x}$ returns the h-margins of $\mathbf{x}$. For this reason, the columns of design matrix W_h are said to form the *marginal basis* for W_h.

Contrast bases and marginal bases are related to each other in a very simple way (see also Theorem 3.7).

Lemma 3.14. $\mathcal{R}(\mathrm{W}_h) = \bigoplus_{h' \colon h' \subseteq h} \mathcal{R}(\mathrm{U}_{h'})$

Proof. The proof is a direct consequence of Equation (19) and part *iii.* of Lemma 3.13. An alternative proof, using directly the properties of the contrasts and marginal design matrices is given below.

Let $\mathrm{B}_h = \bigoplus_{h' \colon h' \not\subseteq h} \mathrm{U}_{h'}$. By Lemma 3.13 part *ii.* and Part *2.* of Lemma 3.13, the column range of B_h spans the null space of $\bigoplus_{h' \colon h' \subseteq h} \mathcal{R}(\mathrm{U}_{h'})$. Therefore, the claim is proved if it is showed that the columns of B_h form a basis for the kernel of $\mathrm{W}_h^\top$. By construction, for every $h' \not\subseteq h$, $\mathrm{U}_{h'}^\top \mathrm{W}_h = 0$, since there exists a $k \in h'$ but $k \notin h$ so that $\mathbf{1}_k^\top \mathrm{U}_k^{h'} = \mathbf{0}$ (see the proof of Lemma 3.13, Part *2.*). Therefore, $\mathrm{W}_h^\top \mathrm{B}_h = 0$. Next, $\mathrm{rank}(\mathrm{W}_h) = \prod_{i \in h} I_k$, which, by Lemma 3.18, is equal to $\sum_{h' \subseteq h} \prod_{k \in h'} (I_k - 1)$, the co-dimension of B_h. Hence $\mathrm{rank}(\mathrm{B}_h) = \dim\left(\mathrm{kernel}(\mathrm{W}_h^\top)\right)$. ■

Using Lemma 3.14 it is possible to compute the dimension of the subspaces $\mathcal{W}_h$ and $\mathcal{V}_h$ (see also Theorem 3.9).

Corollary 3.15. *Let $\mathcal{H}$, Δ be log-linear models and $\mathcal{M}_{\mathcal{H}}, \mathcal{M}_\Delta \subset \mathbb{R}^{\mathcal{I}}$ be the associated subspaces. Then*

$$\dim(\mathcal{M}_{\mathcal{H}}) = \sum_{h \in \mathcal{H}} \prod_{k \in h} (I_k - 1) \tag{30}$$

and

$$\dim(\mathcal{M}_\Delta) = \left(\prod_{k=1}^{K} I_j \right) - \sum_{\{h \in 2^{\mathcal{K}} \colon h \not\subseteq d, \ \ d \in \Delta\}} \prod_{k \in h} (I_k - 1)\,, \tag{31}$$

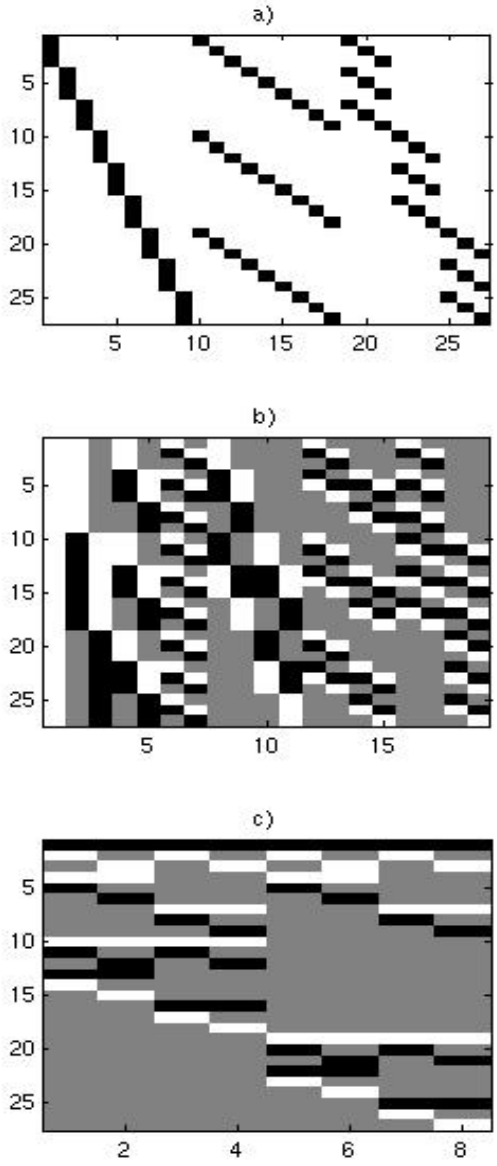
with the convention that, for $h = \emptyset$, $\prod_{k \in h}(I_k - 1) = 1$.

Proof. Equation (30) follows from the orthogonality properties of the design matrices U_h, $h \in \mathcal{H}$, indicated in Lemma 3.13. As for equation (31), since $\mathcal{M}_\Delta = \bigcup_{d \in \Delta} \mathcal{R}(\mathrm{W}_d)$, where d ranges among the facets of the simplicial complex Δ, then, by Lemma 3.14, its orthogonal complement is $\bigcup_{\{h \in 2^{\mathcal{K}} \colon h \notin \Delta\}} \mathcal{R}(\mathrm{U}_h)$. By orthogonality again, the dimension of this subspace is

$$\sum_{h \in 2^{\mathcal{K}} \colon h \notin \Delta} \prod_{k \in h} (I_k - 1)\,,$$

hence the result. ■

Figure 1: 3^3 table and the hierarchical model $\Delta = [12][13][23]$. **a)**: W_Δ design matrix (0 entries in white and $+1$ entries in black). **b)**: U_Δ design matrix (0 entries in gray, -1 entries in back and $+1$ entries in white). **c)**: B_Δ matrix, with columns spanning $\mathcal{M}_\Delta^\perp$ (0 entries in gray, -1 entries in back and $+1$ entries in white).



The proof of the Lemma requires a well-known combinatorial result, given in Section 3.4.3 which uses the Möbius function in exactly the same way as indicated by Theorem 3.9. Equation 31 is given also in Lauritzen (2002), without derivation. A very similar but more involved expression for B_Δ is derived by Hosten and Sullivant (2004) using the notion of *adjacent minors*.

Although the matrices defining the marginal basis are sparse, it is possible to find an even sparser representation of $\mathcal{M}_h$ for hierarchical models. Specifically, let E_n by the first $n-1$ columns of the identity matrix I_n. For any $h \in \mathcal{H}$, let

$$\mathrm{V}_k^h = \begin{cases} \mathrm{E}_k & \text{if} \quad k \in h \\ \mathbf{1}_k & \text{if} \quad k \notin h \end{cases}$$

and

$$\mathrm{V}_h = \bigotimes_{k=1}^{K} \mathrm{V}_k^h.$$

The matrices V_h have the same properties of the matrices U_h listed in Lemma 3.13, even if $\mathcal{R}(\mathrm{V}_h)$ differs from $\mathcal{R}(\mathrm{U}_h)$, for any $h \subset \mathcal{K}$ with $h \neq \emptyset$. However, it is possible to show that

$$\bigoplus_{h' \,:\, h' \subseteq h} \mathcal{R}(\mathrm{V}_{h'}) = \mathcal{R}(\mathrm{W}_h) = \bigoplus_{h' \,:\, h' \subseteq h} \mathcal{R}(\mathrm{U}_{h'}). \tag{32}$$

Similar to the contrast design matrices U_h, efficient ways of generating W_h and V_h row-wise and of storing them in a very compact form are devised and described in Section 3.4.2. These algorithms take advantage of the sparsity of both W_h and V_h as well.

For matrices $\mathrm{A}_1, \ldots, \mathrm{A}_n$ with the same number of rows r and number of columns $c_1, \ldots, c_n$, respectively, we will denote the operation of adjoining them into one matrix of dimension $r \times \sum_k c_k$ with

$$\bigoplus_{k=1}^{n} \mathrm{A}_k = [\mathrm{A}_1 \ldots \mathrm{A}_n] \,.$$

Using this notation, we conclude that

$$\mathrm{U}_\mathcal{H} = \bigoplus_{h \in \mathcal{H}} \mathrm{U}_h$$

is a design matrix for the log-linear model $\mathcal{H}$ and both

$$\mathrm{W}_\Delta = \bigoplus_{d \in \Delta} \mathrm{W}_d \quad \text{and} \quad \mathrm{V}_\Delta = \bigoplus_{d \in \Delta} \mathrm{V}_d \tag{33}$$

are design matrices for the hierarchical log-linear model Δ. In addition, the columns of the matrices

$$\mathrm{B}_\mathcal{H} = \bigoplus_{\{h \in 2^\mathcal{K} \,:\, h \notin \mathcal{H}\}} \mathrm{U}_h$$

and

$$\mathrm{B}_\Delta = \bigoplus_{\{h \in 2^\mathcal{K} \,:\, h \not\subseteq d, \;\; d \in \Delta\}} \mathrm{U}_h \tag{34}$$

form a basis for $\mathcal{M}_\mathcal{H}^\perp$ and $\mathcal{M}_\Delta^\perp$, respectively.

Figure 2: 3^5 table and the random hierarchical model $\Delta = [145][25][135][345][123]$. **a)**: W_Δ design matrix (0 entries in white and $+1$ entries in black). **b)**: U_Δ design matrix (0 entries in gray, -1 entries in back and $+1$ entries in white).

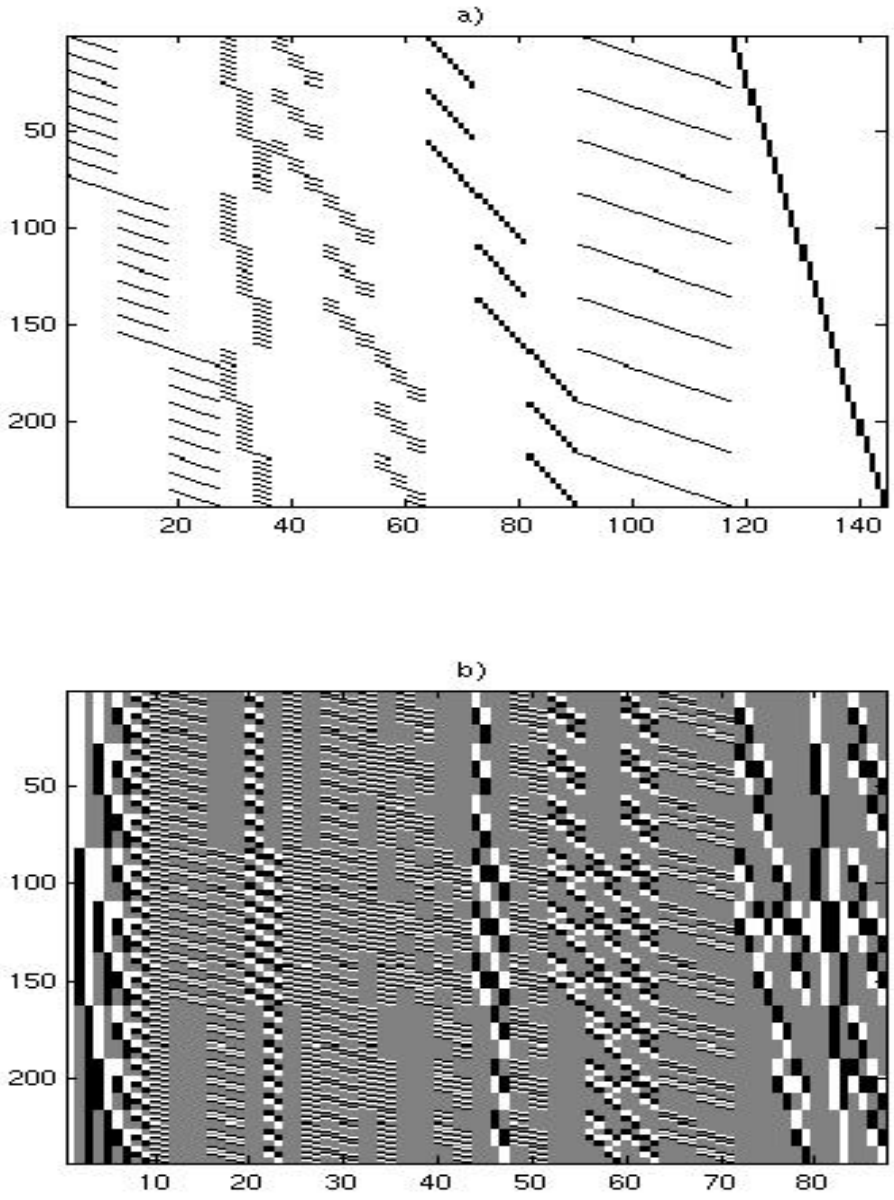
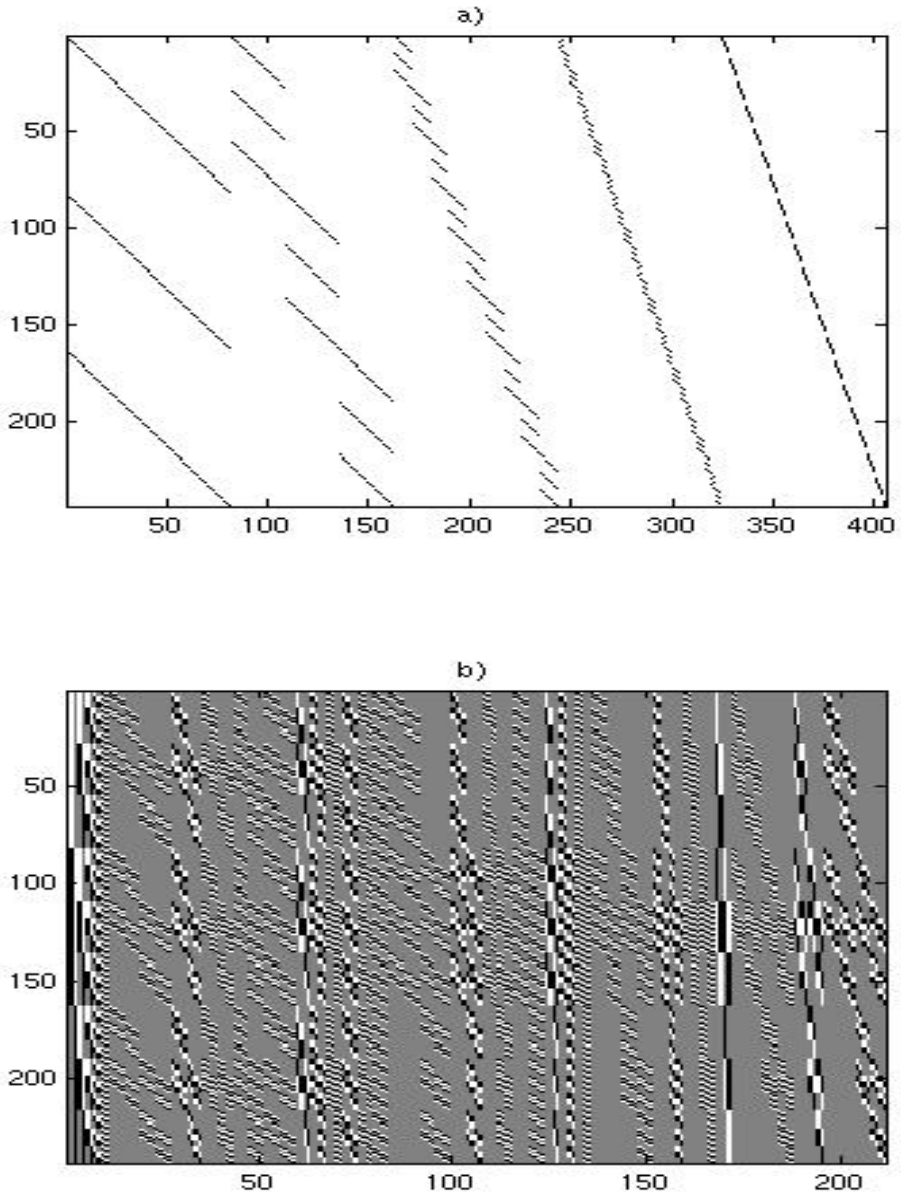


Figure 3: 3^5 table and the hierarchical model $\Delta = [2345][1345][1245][1235][1234]$ (0 entries in white and $+1$ entries in black). **a)**: W_Δ design matrix. **b)**: U_Δ design matrix (0 entries in gray, -1 entries in back and $+1$ entries in white).



Example 3.16 (Design matrices). It is easy to see that, for a log-linear model $\mathcal{H}$, the matrices $\mathrm{U}_{\mathcal{H}}$ and $\mathrm{C}_{\mathcal{H}}$ have

$$\sum_{h\in\mathcal{H}}\prod_{k=1}^{K}\begin{cases} 2(I_k-1) & \text{if} \quad k\in h \\ I_k & \text{if} \quad k\notin h \end{cases}$$

non-zero entries. For a hierarchical model Δ instead, the matrix W_{Δ} has

$$|\Delta|\prod_{k=1}^{K} I_k$$

non-zero entries, where $|\Delta|$ indicates the number of facets of Δ. Figure 1 illustrates some of the matrices introduced in this section for the 3^3 table and the model $\Delta = [12][23][13]$. The 27×27 design matrix W_{Δ} is shown in part **a)** while the full rank 27×19 matrix U_{Δ} is displayed in part **b)**. The matrix W_{Δ} contains 729 entries but only 81 of them are nonzero; in comparison, U_{Δ} has 279 nonzero entries out of 513 total entries. Part **c)** in the Figure shows the full rank 27×8 matrix B_{Δ}, spanning the orthogonal complement of the associated log-linear subspace.

Figure 2 displays the sparse and non-sparse design matrices W_{Δ} and U_{Δ}, respectively, for the 3^5 table and the log-linear model $\Delta = [145][25][135][345][123]$. The sparse matrix has dimension 243×144 and contains only 1215 nonzero entries, while U_{Δ} has smaller dimension 243×87 but is denser, having in fact 8631 nonzero entries. An analogous comparison of the two types of design matrices for the largest hierarchical log-linear model that can be fit to a 5-way table, namely $\Delta = [2345][1345][1245][1235][1234]$, is made in Figure 2. The sparse design matrix W_{Δ} depicted in part **a)** has dimension 243×405 and 1215 nonzero entries, while U_{Δ} has smaller dimension, 243×211, but has more nonzero entries: 15783. ■

3.3 Group Theoretic Derivation

Below we show that the log-linear subspaces presented in this section correspond to the class of log-linear models generated by invariant parametrizations of the underlying probability distribution generating the cell counts with respect to permutations of the cell labels. The basic idea is simple and is based on the observation that the projection matrices of Equation (22) identify irreducible invariant subspaces of $\mathbb{R}^{\mathcal{I}}$. See Serre (1977) for an introduction to group theory and Forster (2003) and Diaconis (1988) and references there within for further details on statistical applications.

Let S_k be the symmetric groups of permutations on the label set $\mathcal{I}_k$ corresponding to the k-th random variable and consider the permutation sub-group on the multi-index $\mathcal{I}$ obtained by composition using the direct product

$$S_{\mathcal{I}} = \prod_{k=1}^{K} S_k.$$

The group $S_{\mathcal{I}}$ consists of all the possible permutations of the label combinations. Letting P_{σ} denote the permutation matrix associated with any permutation $\sigma\in S_{\mathcal{I}}$, it will be shown below that the log-linear models introduced in this section are the linear subspaces $V\subset\mathbb{R}^{\mathcal{I}}$ that are invariant with respect to the permutations in $S_{\mathcal{I}}$: $\mathrm{P}_{\sigma}\mathbf{v}\in\mathrm{V}$ for all $\sigma\in S_{\mathcal{I}}$ and for all $\mathbf{v}\in\mathrm{V}$.

Proposition 3.17. *The irreducible, $S_{\mathcal{I}}$-invariant components of $\mathbb{R}^{\mathcal{I}}$ are the interaction subspaces $\mathcal{U}_h$, $h\in 2^{\mathcal{K}}$, defined in Equation (17).*

Proof. Each symmetric group S_k, $k \in \mathcal{K}$, has $\mathbb{R}^{\mathcal{I}_k}$ as its natural representation space, which in turn possesses two invariant irreducible subspaces, each with multiplicity 1, that is

$$\mathbb{R}^{\mathcal{I}_k} = \mathcal{R}(\mathbf{1}_k) \oplus \mathcal{R}(\mathbf{1}_k)^{\perp},$$

where $\mathbf{1}_k$ denotes the I_k-dimensional vector of ones. By Theorem 10 in Serre (1977), $\mathbb{R}^{\mathcal{I}}$, as a tensor product representation space for $S_{\mathcal{I}}$, has 2^K irreducible invariant components of multiplicity 1:

$$\mathbb{R}^{\mathcal{I}} = \bigotimes_{k=1}^{K} \left(\mathcal{R}(\mathbf{1}_k) \oplus \mathcal{R}(\mathbf{1}_k)^{\perp} \right). \tag{35}$$

Next, for any $h \in 2^{\mathcal{K}}$, equation (27) and Lemma 3.13 imply that

$$\mathcal{R}(\mathrm{U}^h) = \bigotimes_{k \in h} \mathcal{R}(\mathrm{U}_k^h) = \bigotimes_{k \in h} \begin{cases} \mathcal{R}(\mathbf{1}_k) & \text{if} \quad k \in h \\ \mathcal{R}(\mathbf{1}_k)^{\perp} & \text{if} \quad k \notin h, \end{cases}$$

so, that, by pairwise orthogonality of the subspaces $\{\mathcal{R}(\mathrm{U}_h), h \in 2^{\mathcal{K}}\}$ (see Lemma 3.13 again), the decomposition (35) of $\mathbb{R}^{\mathcal{I}}$ into irreducible components becomes

$$\mathbb{R}^{\mathcal{I}} = \bigoplus_{h \subseteq \mathcal{K}} \bigotimes_{k \in h} \mathcal{R}(\mathrm{U}_k^h) = \bigoplus_{h \subseteq \mathcal{K}} \mathcal{R}(\mathrm{U}^h),$$

which coincides with the formula given in part *ii.* of Lemma 3.13. ■

3.4 Appendix

3.4.1 Incorporating Sampling Constraints

In the description so far, it has been assumed lack of sampling constraints (or, equivalently, the Poisson sampling scheme). When sampling is performed according to a non-trivial constraint subspace $\mathcal{N} \subset \mathcal{M}$, special care is needed for dealing with $\mathcal{M} \ominus \mathcal{N}$. In these cases, the log-linear subspaces of this chapter are defined to be the invariant irreducible subspaces of $\mathcal{M} \ominus \mathcal{N}$ rather than $\mathcal{M}$ and, consequently, they may not have a simple representation in general. Fortunately, for product-multinomial sampling with $\mathcal{N} = \mathcal{R}(\mathrm{W}_s)$, where W_s is defined in (29), a representation of the unrestricted log-linear subspace and the corresponding design matrices are readily available. In fact, owning essentially to the orthogonality of the direct sum decomposition of $\mathbb{R}^{\mathcal{I}}$, for hierarchical models the relevant log-linear subspace is

$$\mathcal{M}_{\Delta} \ominus \mathcal{N} = \bigoplus_{\{h:\, h \subseteq d \in \Delta,\ h \not\subseteq s\}} \mathcal{R}(\mathrm{U}^h),$$

while, for general factor-interaction log-linear models, it is given as

$$\mathcal{M}_{\mathcal{H}} \ominus \mathcal{N} = \bigoplus_{\{h \in \mathcal{H},\ h \not\subseteq s\}} \mathcal{R}(\mathrm{U}^h).$$

3.4.2 Generation of U_h, W_h and V_h

Generation of U_h

Assume for convenience and without loss of generality that the elements of h are accessed in increasing order, so that $h = \{k_1, \ldots, k_{|h|}\}$. From the definition of tensor product, it follows that the elements in the $< i_1, \ldots, i_K >$ row of the matrix U_h are

$$u_{<i_1,\ldots,i_K>,<j_1,\ldots,j_K>} = \prod_{k=1}^{K} u^{(k)}_{i_k j_k} = \prod_{k \in h} u^{(k)}_{i_k j_k},$$

where $u^{(k)}_{i_k j_k}$ denotes the (i_k, j_k)-th element of the matrix U^h_k and the index $j_k \in \{1, \ldots, I_k - 1\}$ if $k \in h$ and is equal to 1 otherwise.

Since each row of the matrices U^k_h has at most two non-zero entries, each row of U_h has at most $2^{|h|}$ nonzero elements, which in general will be a minority. It is possible to avoid the repeated computation of zero elements by computing only the nonzero elements and their position in a row as follows. For any $k \in h$, the (i, j)-the entry in the matrix U^h_k is

$$u^{(k)}_{ij} = \begin{cases} -1 & \text{if } j = i - 1 \\ 1 & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a column index j associated with a nonzero element can assume only the values i, if $i < I$ or $i - 1$, if $i > 1$. This suggests that it is possible to code the candidates for nonzero elements of the rows of U_h by a bit string of length $|h|$ whose bits b are numbered $|h|$ though 1. The convention for determining the values of j_{k_b} is

$$j_{k_b} = \begin{cases} i_{k_b} - 1 & \text{if bit } b \text{ is } 0 \\ i_{k_b} & \text{if bit } b \text{ is } 1, \end{cases} \tag{36}$$

for all cases except $(b = 0, i_{k_b} = 1)$ and $(b = 1, i_{k_b} = I_{k_b})$. As the binary value of the bit string ranges from 0 to $2^{|h|} - 1$, the indexes determined by it traverse the row of the basis in the usual lexicographical ordering with the last index varying most rapidly.

For a given bit string, after computing the indexes j_{k_b}, $b = 1, \ldots, |h|$, using (36), the position j of the corresponding possible nonzero element along the row $< i_1, \ldots, i_K >$ of U_h is $< j_{k_1}, \ldots, j_{k_{|h|}} >$ (see equation (1) and recall that the elements of h are ordered in an increasing fashion). These ideas are incorporated in the pseudo-code of Table 8.

If instead the matrix C_k from equation (28) is used to generate of U_h, then, for any $k \in h$, the (i, j)-the entry in the matrix U^h_k is

$$u^{(k)}_{ij} = \begin{cases} 1 & \text{if } i = 1 \\ -1 & \text{if } i = j + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a column index j associated with a nonzero element can take on all the values $1, \ldots, I_k - 1$ if $i = 1$ or only the value $i - 1$, if $i > 1$. In addition, each row of U^h_k has the same sign, a property that consequently holds also for the matrix U_h. The pseudo-code for obtaining the nonzero entries

of a given row of the matrix U_h is given in Table 9.

Generation of W_h
The procedure for generating W_h is very simple since all the entries of any row i of W_h are zeroes except for the entry $< i_h >$ (see Section 1.1 for the notation), which is 1. The algorithm for the row-wise generation of W_h is describe in Table 10.

Generation of V_h
The row of V_h associated to a cell combination $< i_1, \ldots, i_K >$ is 0 if there is a $b \in h$ such that $i_{k_b} = I_{k_b}$. Otherwise, the single one in the row is located at the position

$$\sum_{b=1}^{|h|-1} (i_{k_b} - 1) \left(\prod_{j=b+1}^{|h|} (I_{k_j} - 1) \right) + (i_{k_{|h|}} - 1) + 1.$$

The pseudo-code for the corresponding algorithm is given in Table 11.

3.4.3 A Combinatorial Lemma

Lemma 3.18. *Let $S = \{1, \ldots, n\}$ be a finite set and let $\{d_1, \ldots, d_n\}$ be numbers strictly greater than 1. Then*

$$\left(\prod_{i=1}^{n} d_i \right) - 1 = \sum_{h \subseteq S:\, h \neq \emptyset} \prod_{i \in h} (d_i - 1). \tag{37}$$

Proof. Define the following mappings $\Phi, \Psi : 2^S \to \mathbb{R}_{>0}$ given by $\Phi(h) = \prod_{i \in h}(d_i - 1)$ and $\Psi(h) = \prod_{i \in h} d_i$, with $\Phi(\emptyset) = \Psi(\emptyset) = 1$. Using induction, the following holds true:

$$\Phi(S) = \prod_{i=1}^{n}(d_i - 1) = \sum_{h \subseteq S} (-1)^{|S \setminus h|} \left(\prod_{i \in h} d_i \right) = \sum_{h \subseteq S} (-1)^{|S \setminus h|} \Psi(h). \tag{38}$$

The identity is trivially verified for $|S| = 1$. Assume it is true for any $|S'| = n - 1$ and, without loss of generality, assume $S = S' \cup \{n\}$, with $S' = \{1, \ldots, n-1\}$. Then:

$$\begin{aligned}
\textstyle\prod_{i=1}^{n}(d_i - 1) &= \left(\textstyle\prod_{i=1}^{n-1}(d_i - 1) \right)(d_n - 1) \\
&= \left(\textstyle\sum_{h' \subseteq S'} (-1)^{|S' \setminus h'|} (\prod_{i \in h'} d_i) \right)(d_n - 1) \\
&= \textstyle\sum_{h' \subseteq S'} (-1)^{|S' \cup \{n\} \setminus h' \cup \{n\}|} (\prod_{i \in h' \cup \{n\}} d_i) + \sum_{h' \subseteq S'} (-1)^{|S' \setminus h'| + 1} (\prod_{i \in h'} d_i) \\
&= \textstyle\sum_{h \subseteq S:\, n \in h} (-1)^{|S \setminus h|} (\prod_{i \in h} d_i) + \sum_{h \subseteq S:\, n \notin h} (-1)^{|S \setminus h|} (\prod_{i \in h} d_i) \\
&= \textstyle\sum_{h \subseteq S} (-1)^{|S \setminus h|} \left(\prod_{i \in h} d_i \right),
\end{aligned}$$

proving (38). By Möbius inversion formula (see, for example, Lauritzen, 1996) it follows that

$$\Psi(S) = \sum_{h \subseteq S} \Phi(h).$$

The previous identity, along with $\Phi(\emptyset) = 1$, produces the desired result. ■

4 Computing Extended Maximum Likelihood Estimates

In this section, we derive algorithms for computing extended maximum likelihood estimates. Although the general procedure developed here can be applied to any log-linear subspace, most of the computational considerations and algorithms outlined in the accompanying pseudo-codes are targeted to the types of log-linear models described in Section 3 and, in particular, to hierarchical models. As above, we are only going to concern ourselves with the the Poisson and product-multinomial sampling schemes, although we think that some of the result that follow can be adapted to more general conditional Poisson schemes. Some of the material presented below is based on results due to Fienberg et al. (1980).

Using the language and settings of Section 2, consider the problem of computing the MLE of the cell mean vector for a log-linear model with log-linear subspace $\mathcal{M}$ and sampling subspace $\mathcal{N} \subsetneq \mathcal{M}$, so that there is a total of $k = \dim(\mathcal{M} \ominus \mathcal{N})$ parameters to be estimated. We proved that the MLE is not defined when there is not enough data for estimating completely the cell mean function $\mathbf{m}$, but only the coordinates of $\mathbf{m}$ belonging to the facial set $\mathcal{F} \subset \mathcal{I}$ corresponding to the observed sufficient statistics. Facial sets depend on the observed data and on the geometric and combinatorial structure of the polyhedral cone generated by the rows of the design matrix U, where $\mathcal{R}(\mathrm{U}^\top) = \mathcal{M} \ominus \mathcal{N}$. For each the facial set $\mathcal{F}$ there is an associated k'-dimensional subspace $\mathcal{M}_\mathcal{F}$ of $\mathbb{R}^\mathcal{F}$, such that $\log \mathbf{m}_\mathcal{F} \in \mathcal{M}_\mathcal{F}$. In fact, only k' of the k original parameters can be estimated. The subspace $\mathcal{M}_\mathcal{F}$ is completely determined by the facial set $\mathcal{F}$ in the sense that $\mathcal{M}_\mathcal{F}$ is spanned by the rows of $\mathrm{U}_\mathcal{F}$, where $\mathrm{U}_\mathcal{F}$ denotes the matrix derived from U by considering only the columns in $\mathcal{F}$. From the viewpoint of the natural parameter space, nonexistence of the MLE implies that there is enough data to estimate parameters only along a k'-dimensional flat of $\mathbb{R}^k$. Such a hyperplane is an affine transform of $\mathbb{R}^{k'}$, which can be taken to be the natural parameter space of the restricted, minimally-represented exponential family of the log-linear model determine by $\mathcal{M}_\mathcal{F}$. See Rinaldo (2006) for details. Provided the log-likelihood of the restricted family is parametrized in minimal form, it will be a strictly concave function on $\mathbb{R}^{k'}$, admitting a unique optimum, which corresponds to the extended MLE. Once the facial set $\mathcal{F}$ is available, this amounts to isolating any set of linearly independent rows of $\mathrm{U}_\mathcal{F}$ and using them to re-parametrize the restricted log-likelihood function.

The computation of the MLE and extended MLE proceeds in two fundamental steps. The inputs of this procedure are the design matrix U and the log-likelihood function $\ell : \mathbb{R}^k \to \mathbb{R}$, taking as argument the k-dimensional vector (of natural parameters) $\boldsymbol{\theta}$ such that $\boldsymbol{\mu} = \mathrm{U}^\top \boldsymbol{\theta} \in \mathcal{M}$.

1. **The determination of the facial set (Section 4.1).**
   Computing the facial set is a task that can be described as in generality:

   Given a conic integer combination $\mathbf{z}$ *of the columns of* U*, determine the set* $\mathcal{F}$ *of those columns which belong to the face of the associated polyhedral cone containing* $\mathbf{z}$ *in its relative interior.*

   For this task, the design matrix does not have to be of full rank and its column range can either be $\mathcal{M} \ominus \mathcal{N}$ or $\mathcal{M}$. From the computation viewpoint, however, design matrices of full-row rank and integer entries are preferable.

2. **The maximization of the possibly restricted log-likelihood function (Section 4.3).**
   After obtaining the appropriate facial set $\mathcal{F}$, if $\mathcal{F} \subsetneq \mathcal{I}$ a new, reduced design matrix of full row rank U^* is computed by selecting any subset of linearly independent rows from $\mathrm{U}_\mathcal{F}$, as described in Section 4.4. The log-likelihood function is re-parametrized by U^* and consequently

is re-defined as a function $\ell^* : \mathbb{R}^{k'} \to \mathbb{R}$, where $k' = \mathrm{rank}(\mathrm{U}^*) < k$. The Newton-Raphson procedure is used to maximize ℓ or ℓ^*, depending whether the MLE exists or not. Using full-dimensional design matrices U or U^*, which is *de facto* equivalent to representing the corresponding linear exponential families in minimal form, guarantees that the log-likelihood functions ℓ or ℓ^*, respectively, are strictly concave on their entire domain, and hence admits a unique (and finite, in this case) maximizer.

The outcome of this procedure is the extended MLE of the cell mean vector, whose support is the relevant facial set $\mathcal{F}$. When the MLE is non-existent, an important by-product of step 2. above is a basis for the subspace $\mathcal{M}_{\mathcal{F}}$, whose dimension is also the dimension of the boundary log-linear model, or the order of the reduced exponential family corresponding to $\mathcal{F}$ (see Rinaldo, 2006). This information is crucial for computing the correct number of degrees of freedom for the asymptotic χ^2 approximation to various measures of goodness of fit and, consequently, for performing goodness-of-fit testing model selection in a correct fashion, as described in Section 6.

In the reminder of the section we will provides an array of different techniques for identifying facial sets and for maximizing the log-likelihood functions under Poisson and product-multinomial sampling schemes. Although the procedures we are about to present are correct in theory, a further computational investigation for ascertaining their efficiencies is in order, preferably focusing primarily on the case of large and sparse datasets. Some of the procedures described below were implemented in a small `MATLAB` toolbox. As these routines were written primarily for testing purposes, they are limited to the Poisson scheme only and, in most cases, do not take advantage of the proposed algorithms for minimizing memory usage and computational complexity. Nevertheless, they are functional, rather easy to work with and can be used to compute both the MLE and extended MLE. The toolbox is available on-line at `www.stat.cmu.edu/~arinaldo/ExtMLE/` and was written using `MATLAB` version `7.0.4`.

4.1 Determination of the Facial Sets

In this section, we present and discuss two methods for determining facial sets, one based on linear programming and the other on the maximization of a well-behaved non-linear function via Newton-Raphson procedure. Alternative methodologies, still of theoretical interest but perhaps of less practical use, are described in Appendix A. The applicability of the procedures developed here is quite general, as they do not rely on any specific assumption about the sampling scheme utilized or about the type of log-linear models considered, although they are particularly efficient under Poisson and product-multinomial schemes and when the design matrices are computed using the methods described in Section 3. We will describe here two procedure

For convenience we will now work with the transpose of the design matrices we have been considering so far and, specifically, we let U be the design matrix whose rows are indexed by elements in the cell set $\mathcal{I}$. Denote with U_+ and U_0 the sub-matrices obtained from U by considering the rows in $\mathcal{I}_+ := \mathrm{supp}(\mathbf{n})$ and $\mathcal{I}_0 := \mathrm{supp}(\mathbf{n})^c$, respectively. Recall that each face F of the marginal cone $\mathrm{cone}(\mathrm{U}^\top)$ is uniquely identified by the associated facial set $\mathcal{F} \subset \mathcal{I}$ such that, for some vector $\boldsymbol{\zeta}_F$,

$$\begin{cases} \mathbf{u}_i^\top \boldsymbol{\zeta}_F = 0 & \text{if} \quad i \in \mathcal{F} \\ \mathbf{u}_i^\top \boldsymbol{\zeta}_F > 0 & \text{if} \quad i \in \mathcal{F}^c \end{cases}, \tag{39}$$

where $\mathbf{u}_i$ denotes the i-th row of U and the set $\mathcal{F}^c = \mathcal{I} \setminus \mathcal{F}$ will be called *co-facial set*.

Equation (39) implies that the observed sufficient statistics $\mathbf{t} = \mathrm{U}^\top \mathbf{n}$ belong to the relative interior of some proper face F of the marginal cone if and only if the associated co-facial set $\mathcal{F}^c$ satisfies the inclusion $\mathcal{F}^c \subseteq \mathcal{I}_0$. This, in turn, is equivalent to the existence of a vector $\mathbf{c}^*$ satisfying:

1. $\mathrm{U}_+\mathbf{c}^* = \mathbf{0}$;
2. $\mathrm{U}_0\mathbf{c}^* \gtrsim \mathbf{0}$;
3. the non empty set $\mathrm{supp}(\mathrm{U}\mathbf{c}^*)$ has maximal cardinality among all support vectors of the type $\mathrm{supp}(\mathrm{U}\mathbf{x})$ with $\mathrm{U}\mathbf{x} \gtrsim \mathbf{0}$.

Therefore, any solution of the non-linear optimization problem

$$\begin{array}{ll} \max & |\mathrm{supp}(\mathrm{U}\mathbf{x})| \\ \text{s.t.} & \mathrm{U}_+\mathbf{x} = 0 \\ & \mathrm{U}_0\mathbf{x} \geq \mathbf{0} \end{array} \tag{40}$$

will identify the required co-facial set $\mathcal{F}^c = \mathrm{supp}(\mathrm{U}\mathbf{x}^*)$.

The problem (40) can be simplified making use of the following fact.

Lemma 4.1. *The MLE exists if* $\mathrm{rank}(\mathrm{U}_+) = \mathrm{rank}(\mathrm{U})$.

Proof. The stated condition follows immediately from this other restatement of Theorem 2.3 in Haberman (1974):
The MLE exists if and only if there exists a vector $\mathbf{y} > \mathbf{0}$ *such that* $\mathrm{U}_0^\top \mathbf{y} \in \mathcal{R}(\mathrm{U}_+^\top)$.

This claim follows from an application of Motzkin's Transposition Theorem (Schrijver, 1998, page 94) to the conditions given in Theorem 2.3 in Haberman (1974). An alternative proof is the following. By Theorem 2.2 in Haberman (1974), the MLE exists if and only if there exists a $\boldsymbol{\delta} \in \mathrm{kernel}(\mathrm{U}^\top)$ with $\boldsymbol{\delta}_{\mathcal{I}_0} > \mathbf{0}$. This occurs if and only if $\mathrm{U}_+^\top \mathbf{n}_{\mathcal{I}_+} = \mathrm{U}_+^\top(\mathbf{n}_{\mathcal{I}_+} + \boldsymbol{\delta}_{\mathcal{I}_+}) + \mathrm{U}_0^\top \boldsymbol{\delta}_{\mathcal{I}_0}$, which is equivalent to $\mathrm{U}_0^\top \boldsymbol{\delta}_{\mathcal{I}_0} = \mathrm{U}_+^\top(-\boldsymbol{\delta}_{\mathcal{I}_+})$. The result follows. ■

Example 4.2. The condition of Lemma 4.1 is only sufficient. As a counter-example, consider the 3-way table

0	0		0	0				
0		0	0	0				
								0

and the model $\Delta = [12][13][23]$. It can be verified that the MLE is defined but $\mathrm{rank}(\mathrm{U}_+) = 18$ and $\mathrm{rank}(\mathrm{U}) = 19$. ■

Proposition 4.1 says that it is necessary to look for a facial set only when $\mathrm{rank}(\mathrm{U}) > \mathrm{rank}(\mathrm{U}_+)$. Then, if this is in fact the case, consider the matrix $\mathrm{A} = \mathrm{U}_0\mathrm{Z}$, where the columns of Z form a basis for $\mathrm{kernel}(\mathrm{U}_+)$. Next, observe that (the permutation of the elements of) any vector $\mathbf{y} \in \mathcal{R}(\mathrm{U})$ with $\mathbf{y}_{\mathcal{I}_0} = \mathbf{0}$ can be written as

$$\mathbf{y} = \mathrm{UZ}\mathbf{x} = \begin{pmatrix} \mathrm{U}_+\mathrm{Z}\mathbf{x} \\ \mathrm{U}_0\mathrm{Z}\mathbf{x} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathrm{A}\mathbf{x} \end{pmatrix},$$

for some $\mathbf{x} \in \mathbb{R}^q$ with $q = \mathrm{codim}(\mathcal{R}(\mathrm{U}_+))$. Then, another condition for existence of the MLE follows readily.

Corollary 4.3. *The MLE exists if and only if the system* $\mathrm{A}\mathbf{x} \gneq \mathbf{0}$ *if infeasible.*

Remark.
Using the language of matroids, deciding whether the MLE exists or not is equivalent to the task of deciding whether the rows of A form a totally cyclic vector configuration (see Ziegler, 1998, Chapter 6). This idea will be developed later in Section 4.5.3.

In virtue of Corollary cor:altern, the problem (40) can then take on the simpler form

$$\begin{array}{ll} \max & |\mathrm{supp}(\mathrm{A}\mathbf{x})| \\ \text{s.t.} & \mathrm{A}\mathbf{x} \geq 0. \end{array} \tag{41}$$

The nonzero rows of A are indexed in a natural way by the corresponding subset of $\mathcal{I}_0$, denoted $\mathcal{I}_{\mathrm{A}}$. In the remainder of the section it is assumed, without loss of generality, that A does not have any zero rows, namely $\mathrm{A} = \mathrm{A}_{\mathcal{I}_{\mathrm{A}}}$. As above, any optimal solution $\mathbf{x}^*$ of (41) will provide the co-facial set $\mathcal{F}^c = \mathrm{supp}(\mathrm{A}\mathbf{x}^*)$.

In order to compute the matrix A, a basis for $\mathrm{kernel}(\mathrm{U}_+)$, if different than the trivial subspace $\{\mathbf{0}\}$, must be computed. This can be accomplished using one the methods discussed in Section 4.4 through Equation (62). In addition, for those cases based on product-multinomial sampling for which it is desirable to work with the reduced subspace $\mathcal{M} \ominus \mathcal{N}$, the procedure is modified as follows. Arguing as in later Section 4.3.5, let U be partitioned in the form $\mathrm{U} = (\mathrm{U}_1, \mathrm{U}_2)$, where the columns of U_1 form a basis for the sampling subspace spanned by $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r$. Set

$$\mathrm{V} = \mathrm{U}_2 - \mathrm{U}_1\mathrm{W}_+,$$

where $\mathrm{W}_+ = \mathrm{D}_+^{-1}(\mathrm{U}_+^{(1)})^\top \mathrm{U}_+^{(2)}$, with $\mathrm{D}_+ = (\mathrm{U}_+^{(1)})^\top \mathrm{U}_+^{(1)}$ diagonal and non-singular. It can be seen that the columns of the matrix $(\mathrm{U}_1, \mathrm{V})$ span $\mathcal{M}$. Moreover, by construction, the columns of V_+ are orthogonal to the columns of U_1 (see Equation (59)). It follows from the independence of the columns of U_1 that any basis for the null space of $(\mathrm{U}_1, \mathrm{V})_+$ must be of the form

$$\begin{pmatrix} 0 \\ \mathrm{Z} \end{pmatrix},$$

i.e. the entire dependency resides in the columns of V. Then the matrix

$$\mathrm{A} = (\mathrm{U}_1, \mathrm{V})_+ \begin{pmatrix} 0 \\ \mathrm{Z} \end{pmatrix} = \mathrm{V}_+\mathrm{Z}$$

can be used for determining the facial set by solving the problem (41), previous elimination of possible redundant zero rows.

In the reminder of the section, two methods for finding a solution to problem (41) will be discussed. The appendix contains other proposed procedures.

4.1.1 Linear Programming

Although the problem (41) is highly non-linear, linear programming (LP) methods can still be used to compute its solution. The non-linearity is in fact problematic to the extent that it will typically

require repeated implementations of LP algorithms, whose complexity, however, decreases at each iteration.

In order to describe the basic idea behind the linear programming approximation, let $\mathrm{C_U} = \mathrm{cone}(\mathrm{U}^\top)$ be the marginal cone whose dual is

$$\widehat{\mathrm{C}}_\mathrm{U} = \{\mathbf{x}\colon \mathbf{y}^\top\mathbf{x} \geq 0, \forall \mathbf{y} \in \mathrm{C_U}\},$$

so that the face lattice of $\mathrm{C_U}$ is the opposite of the face lattice of $\widehat{\mathrm{C}}_\mathrm{U}$. Equivalently, the co-facial set of $\mathrm{C_U}$ are the facial sets of $\widehat{\mathrm{C}}_\mathrm{U}$, and vice versa.

A linear version of problem (40) is the linear program

$$\begin{array}{ll} \max & \left(\mathbf{1}_0^\top \mathrm{U}_0\right)\mathbf{x} \\ \text{s.t.} & \mathrm{U}_+\mathbf{x} = 0 \\ & \mathrm{U}_0\mathbf{x} \geq \mathbf{0} \\ & \mathrm{U}_0\mathbf{x} \leq \mathbf{1}, \end{array} \tag{42}$$

where the third constraint is required to bound the value of the objective function. The feasible set contains $\mathrm{kernel}(\mathrm{U})$ and is contained in $\widehat{\mathrm{C}}_\mathrm{U}$. If $\mathbf{x} \in \mathrm{kernel}(\mathrm{U})$, the objective function takes on its maximum value 0. In fact the MLE exists if and only if the feasible set reduces to $\mathrm{kernel}(\mathrm{U})$. If the MLE does not exist, the vector $\left(\mathbf{1}_0^\top \mathrm{U}_0\right)$ is normal to the supporting hyperplane (in $\mathbb{R}^k$) of some face of $\hat{\mathrm{C}}_\mathrm{U}$ dual to some face containing the observed margins. If a positive optimum is found at some point $\mathbf{x}^*$, then $\mathrm{supp}(\mathrm{U}_0\mathbf{x}^*)$ gives a co-face corresponding to a face on which $\mathbf{t}$ lies. However, there is no guarantee that $\mathbf{x}^*$ is such that $\mathrm{U}_0\mathbf{x}^*$ has maximal support, given the constraints. Geometrically, this means that the procedure is not guaranteed to produce exactly the face whose relative interior contains $\mathbf{t}$, as it might very well produce a face whose relative boundary contains $\mathbf{t}$ (see Example 4.4). The only case in which the LP problem (42) provides the appropriate face with certainty is when $\mathbf{t}$ lies on the relative interior of a facet of $\mathrm{C_U}$, because the correspondence in the dual cone is given by an extreme ray. This fact was exploited in the polynomial time algorithm given in Eriksson et al. (2006).

It is convenient to take advantage of the simplified problem (41) to re-formulate (42) more compactly and efficiently. The corresponding linear program is

$$\begin{array}{rccc} & \max \mathbf{1}^\top\mathbf{y} & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{A}\mathbf{x} \\ & \mathbf{y} & \geq & \mathbf{0} \\ & \mathbf{y} & \leq & \mathbf{1}. \end{array} \tag{43}$$

The set $\mathrm{DC_A} := \{\mathbf{y} \geq \mathbf{0}\colon \mathbf{y} = \mathrm{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^k\}$ is a polyhedral cone which is isomorphic (and combinatorially equivalent) to the polyhedral cone $F_0 := \{\mathbf{x} \in \widehat{\mathrm{C}}_\mathrm{U}\colon \mathrm{U}_+\mathbf{x} = \mathbf{0}\}$, which is precisely the face of $\widehat{\mathrm{C}}_\mathrm{U}$ dual to the face containing $\mathbf{t}$ in its relative interior.

Letting C_0 be the intersection of the unit-hypercube and the non-negative orthant in $\mathbb{R}^{\mathcal{I}_\mathrm{A}}$, the polytope $\mathrm{DC}_\mathrm{A}^{[\mathbf{0},\mathbf{1}]} = \mathrm{DC_A} \cap C_0$ consists of the set of points satisfying the constraints in equation (43). At the optimum $(\mathbf{x}^*, \mathbf{y}^*)$, $\mathbf{1}'\mathbf{y}^* = 0$ if and only if the MLE exist, which happens if and only if $\mathrm{DC}_\mathrm{A}^{[\mathbf{0},\mathbf{1}]} = \{\mathbf{0}\}$. When the MLE does not exist, then for each $\mathbf{y}$ in $\mathrm{DC}_\mathrm{A}^{[\mathbf{0},\mathbf{1}]}$, $\mathrm{supp}(\mathbf{y}) \subseteq \mathcal{F}^c$, with equality if and only if $\mathcal{I}_\mathrm{A} = \mathcal{F}^c$. Then, the procedure (41) aims at finding a point in $\mathrm{DC}_\mathrm{A}^{[\mathbf{0},\mathbf{1}]}$ with maximal support (which does not necessarily have to lie in the relative interior of $\mathrm{DC}_\mathrm{A}^{[\mathbf{0},\mathbf{1}]}$). As indicated above and demonstrated in the following example, one round of the program (43) may not be sufficient to provide the appropriate facial set.

Example 4.4. For the case of 4^3 tables and the model of no-3-factor effect $\Delta = [12][23][13]$, consider the pattern of likelihood zeros

0	0	0	0	0			0	0	0	0		0		0	
	0	0			0		0	0	0	0	0		0	0	
		0	0	0	0	0	0					0		0	0
0		0	0		0		0	0	0	0	0				0

obtained by taking the union of two among the 113,740 possible patterns of likelihood zeros characterizing the facets of the corresponding marginal cone (see Table 1 in Eriksson et al., 2006). Using the MATLAB routine `linprog`[1], it was observed that one application of the LP procedure identifies only a subset of likelihood zeros, namely

	0		0				0		0						
							0	0			0				
		0	0			0									
									0						

and that the complete patterns is correctly determined using a second iteration, after removing the likelihood zeros found in the first one. ■

As the previous example suggests, repeated applications of (43) will eventually produce the required co-face: replace A with $\mathrm{A}_{\mathrm{supp}(\mathrm{A}\mathbf{x}^*)^c}$ at each step until either the objective function is 0 or $\mathrm{supp}(\mathrm{A}\mathbf{x}^*)^c = \emptyset$. The pseudo-code for the LP procedure is given in Table 1.

```
1       : F = I
2       : do repeat
2.1     :     Compute a solution (y*, x*) of (43).
2.2     :     if 1^T y* = 0
2.2.1   :         return F
2.3     :     else
2.3.1   :         F = F \ supp(Ax*)
2.3.2   :         if supp(Ax*)^c = ∅
2.3.2.1 :             return F
2.3.3   :         else
2.3.3.1 :             A = A_{supp(Ax*)^c}
2.3.4   :         end
2.4     :     end
3       : end
```

Table 1: Pseudo-code for the LP procedure to compute the facial set $\mathcal{F}$.

[1]The default optimization options for `linprog` were used: `options=optimset('Simplex','off','LargeScale','on')`.

4.1.2 Newton-Raphson Procedure

In this section, we describe a non-linear optimization problem whose solution will also solve 41. This problem can be attacked using the Newton-Raphson method, is guaranteed to produce the appropriate facial set and, unlike the LP method presented above, does not need repetitions.

Let the function $f\colon \mathbb{R}^k \to \mathbb{R}$ be defined as

$$f(\mathbf{x}) = -\mathbf{1}^\top \exp(\mathrm{A}\mathbf{x}), \tag{44}$$

with gradient $\nabla f(\mathbf{x}) = -\mathrm{A}^\top \exp^{\mathrm{A}\mathbf{x}}$ and hessian $\nabla^2 f(\mathbf{x} =) - \mathrm{A}^\top \exp^{\mathrm{A}\mathbf{x}} \mathrm{A}$, which is negative definite for each $\mathbf{x} \in \mathbb{R}^k$. The following proposition relates the problem of optimizing f with the existence of the MLE. In particular, we show that the existence of a facial set is signaled by a diverging behavior of the Newton sequence $\{\mathbf{x}_k\}$ (see Section 4.6). In addition, when the MLE is nonexistent, the sequence of points $\{\mathbf{x}_n\}$ realizing the supremum of f is not only diverging, but it is guaranteed to eventually identify the appropriate co-face.

Proposition 4.5. *Let f be as in (44) and consider the optimization problem*

$$\sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x}). \tag{45}$$

The MLE exists if and only if the unique optimum of the problem (45) is attained for a finite vector $\mathbf{x}^* \in \mathbb{R}^k$. If the MLE does not exist, $\mathrm{supp}(\lim_n \exp^{\mathrm{A}\mathbf{x}_n^*})^c = \mathcal{F}^c$.

Proof. It easy to see that the function $f(\mathbf{x})$ is bounded from above and strictly concave on $\mathbb{R}^k$. Suppose the unique optimum is attained for some vector $\mathbf{x}^* \in \mathbb{R}^k$, so that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Letting $\mathbf{y}^* = \exp^{\mathrm{A}\mathbf{x}^*} > \mathbf{0}$, the optimality condition on the gradient implies that $\mathrm{A}^\top \mathbf{y}^* = \mathbf{0}$. By Stiemke's Theorem 4.19, the system $\mathrm{A}\mathbf{x} \gneq \mathbf{0}$ has no solutions hence the MLE exists. To show the converse, suppose the MLE does not exist but the optimum in (45) is attained for a finite vector $\mathbf{x}^*$. Then, there exists a subset (possibly improper) $\mathcal{F}^c$ of the row indices $\mathcal{I}_\mathrm{A}$ and a sequence $\{\mathbf{w}\}_n$ such that $\mathbf{a}_i^\top \mathbf{w}_n < 0$ for each n and $\mathbf{a}_i^\top \mathbf{w}_n \downarrow -\infty$ if $i \in \mathcal{F}^c$, while $\mathbf{a}_i^\top \mathbf{w}_n = 0$ for each n if $i \notin \mathcal{F}^c$. It follows that $f(\mathbf{x}^* + \mathbf{w}_n)$ is increasing in n and strictly bigger than $f(\mathbf{x}^*)$ for all n, which gives a contradiction. Hence the optimum is achieved in the limit for a sequence of points $\{\mathbf{x}_n^*\}$ with $||\mathbf{x}_n^*|| \to \infty$ such that $\lim_n f(\mathbf{x}_n^*) = \sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x})$. See also Borwein and Lewis (2000, Theorem 2.2.6).

To prove the last statement, let $\{\mathbf{x}_n^*\}$ be a sequence such that $\lim_n f(\mathbf{x}_n^*) = \sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x})$ and let $\mathbf{y}^* = \lim_n \exp^{\mathrm{A}\mathbf{x}_n^*}$. It is clear that $\mathrm{supp}(\mathbf{y}^*)^c \subseteq \mathcal{F}^c$. In fact, for any $i \in \mathrm{supp}(\mathbf{y}^*)^c$, there exists a subsequence $\{\mathbf{x}_{n_k}^*\}$ such that, eventually, $\mathbf{a}_i^\top \mathbf{x}_{n_k}^* < 0$ for all n_k big enough. This implies that $i \in \mathcal{F}^c$. To show the opposite inclusion $\mathcal{F}^c \subseteq \mathrm{supp}(\mathbf{y}^*)^c$ suppose there exists a $i \in \mathcal{F}^c$ which does not belong to $\mathrm{supp}(\mathbf{y}^*)^c$. Then, letting $\{\mathbf{w}\}_n$ be defined as above, $\mathbf{a}_i^\top \mathbf{w}_n \downarrow -\infty$ but $\lim_n |\mathbf{a}_i^\top \mathbf{x}_n^*| < \infty$, so that

$$\lim_n f(\mathbf{x}_n^* + \mathbf{w}_n) > \lim_n f(\mathbf{x}_n^*) = \sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x}),$$

a contradiction. ■

Remark.
If the MLE does not exist and $\mathcal{I}_\mathrm{A} = \mathcal{F}^c$, then $\sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x}) = 0$.

As already mentioned, the function (44) can be optimized using Newton-Raphson method. In fact, it can be shown that, when the MLE exists, (44) satisfies the assumptions of Section 4.6 over

any neighborhood of the solution $\mathbf{x}^*$, so that quadratic convergence is achieved. When the MLE does not exists, the Newton-Raphson procedure can still be applied and, in addition, it will return the correct facial set by producing a divergent Newton sequence. The result that ensures this divergence is contained in the following Theorem, which shows that the Newton sequence will in fact produce a sequence of points with exploding norm, since the Newton step will always increase in the objective function.

Theorem 4.6. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a strictly concave function of class $\mathcal{C}^3$, strongly concave on any bounded ball and having no maximum on the closure of the open ball B. For any $\mathbf{x} \in B$ let $\mathbf{d}$ be the Newton direction corresponding to $\mathbf{x}$. Then, there exists a positive constant γ independent of $\mathbf{x}$ and a positive number $\alpha \leq 1$ such that:*

$$f(\mathbf{x} + \alpha\mathbf{d}) - f(\mathbf{x}) \geq \gamma. \tag{46}$$

Proof. Let B' be the smallest ball containing the bounded set:

$$\{\mathbf{x} + \mathbf{d}_\mathbf{x}, : \mathbf{x} \in B, \mathbf{d}_\mathbf{x} = -\nabla^2 f(\mathbf{x})\nabla f(\mathbf{x})\}$$

so that $B \subset B'$. Using strict concavity and uniform concavity on B' of f, there exist positive constants K and $L > 1$ such that

$$K \leq -\mathbf{y}^\top \nabla^2 f(\mathbf{x})^{-1}\mathbf{y} \leq L \tag{47}$$

for all $\mathbf{x} \in B$ and all unit vectors $\mathbf{y}$ (i.e. $||\mathbf{y}|| := \mathbf{y}^\top\mathbf{y} = 1$). Since

$$\mathbf{d} = -\nabla^2 f(\mathbf{x})^{-1}\nabla f(\mathbf{x})$$

it follows that

$$K||\nabla f(\mathbf{x})|| \leq ||\mathbf{d}|| \leq L||\nabla f(\mathbf{x})||. \tag{48}$$

Using Taylor's expansion,

$$f(\mathbf{x} + \alpha\mathbf{d}) = f(\mathbf{x}) + \alpha\nabla f(\mathbf{x})^\top\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^\top\nabla^2 f(\mathbf{x} + c\alpha\mathbf{d})\mathbf{d}, \tag{49}$$

for some $0 < c < 1$. Next, it is possible to bound the right hand side of (49), by taking advantage of (47) and (48). In fact,

$$\begin{aligned} \alpha\nabla f(\mathbf{x})^\top\mathbf{d} &= -\alpha\tfrac{\nabla f(\mathbf{x})^\top}{\|\nabla f(\mathbf{x})\|}\nabla^2 f(\mathbf{x})\tfrac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\|\nabla f(\mathbf{x})\|^2 \\ &\geq \alpha K\|\nabla f(\mathbf{x})\|^2 \end{aligned}$$

and

$$\begin{aligned} \tfrac{\alpha^2}{2}\mathbf{d}^\top\nabla^2 f(\mathbf{x} + c\alpha\mathbf{d})\mathbf{d} &= -\tfrac{\alpha^2}{2} - \tfrac{\mathbf{d}^\top}{\|\mathbf{d}\|}\nabla^2 f(\mathbf{x} + c\alpha\mathbf{d})\tfrac{\mathbf{d}}{\|\mathbf{d}\|}\|\mathbf{d}\|^2 \\ &\geq -\tfrac{\alpha^2}{2}L\|\mathbf{d}\|^2 \\ &\geq -\tfrac{\alpha^2}{2}L^2\|\nabla f(\mathbf{x})\|^2. \end{aligned}$$

Therefore,

$$f(\mathbf{x} + \alpha\mathbf{d}) - f(\mathbf{x}) \geq \left(\alpha K - \frac{\alpha^2}{2}L^2\right)\|\nabla f(\mathbf{x})\|^2.$$

Since $\sup_{0\leq\alpha\leq 1}\left(\alpha K - \frac{\alpha^2}{2}L^2\right) = \frac{1}{2}\frac{K^2}{L^2} > 0$, choose $0 \leq \alpha' \leq 1$ so that $\tau := \left(\alpha' K - \frac{(\alpha')^2}{2}L^2\right) > 0$ and let $\gamma = \tau\left(\inf_{\mathbf{x}\in\bar{B}}||\nabla f(\mathbf{x})||^2\right) > 0$, where the last inequality holds since f has no maximum on the closure of B. Then, for such a choice of α and γ, $f(\mathbf{x} + \alpha\mathbf{d}) - f(\mathbf{x}) \geq \gamma$, as desired. ■

4.2 Existence of the MLE and Markov Bases

Existence of the MLE can be established if the knowledge of the Markov basis (see Diaconis and Sturmfels, 1998) associated to a generic design matrix X is available. The results presented below are more of theoretical interest, since Markov bases can only be computed for small models and tend to contains a large number of elements.

Let $\mathcal{L} = \text{kernel}(\mathrm{X}^\top) \cap \mathbb{Z}^{\mathcal{I}}$. A geometric object which has relevance in statistics is the *fiber*, the set of all contingency tables whose margins match the observed one. Points in the fiber are points in the support of the conditional distribution of the counts given the sufficient statistics, often known as the "exact distribution".

Definition 4.7. Given a point $\mathbf{u} \in \mathbb{N}^{\mathcal{I}}$, the *fiber* at $\mathbf{u}$, denoted by $\mathcal{Z}(\mathbf{u})$, is the congruence class of $\mathbf{u}$ modulo $\mathcal{L}$, i.e.

$$\mathcal{Z}(\mathbf{u}) = \{\mathbf{v} \in \mathbb{N}^{\mathcal{I}} \ : \ \mathbf{u} - \mathbf{v} \in \mathcal{L}\}.$$

Let $\mathcal{B} \subseteq \mathcal{L}$ and, for each $\mathbf{u} \in \mathbb{N}^{\mathcal{I}}$, let $\mathcal{Z}_{\mathcal{B}}(\mathbf{u})$ be the undirected graph whose nodes are the elements of the fiber $\mathcal{Z}(\mathbf{u})$ and in which two nodes $\mathbf{v}$ and $\mathbf{v}'$ are connected if and only if either $\mathbf{v} - \mathbf{v}' \in \mathcal{B}$ or $\mathbf{v}' - \mathbf{v} \in \mathcal{B}$.

Definition 4.8. A set $\mathcal{B} \subset \mathcal{L}$ is said to be a Markov basis if $\mathcal{Z}_{\mathcal{B}}(\mathbf{u})$ is connected for each $\mathbf{u} \in \mathbb{N}^{\mathcal{I}}$.

Although Markov bases are by no means unique, for a given design matrix, all Markov bases which are minimal with respect to inclusion have the same cardinality (see Takemura and Aoki, 2004), so that it is customary to talk about "the" Markov Basis.

If a Markov Basis is available, it is possible to detect whether the observed margins $\mathbf{t}$ lie on a face of the marginal cone and, in this case, even to identify such a face. In fact, Markov bases are rich enough to fully characterize the combinatorial structure of the marginal cone and hence can be used to identify facial sets. These facts are proved in the next theorem and in the subsequent corollary and suggest a simple algorithm for obtaining facial sets sketched in Table 2.

Denoting with M the $\mathcal{I} \times m$ matrix whose columns are the elements of the (minimal) Markov Basis for $\mathcal{L}$, let $\mathrm{M}^\sigma = \text{sgn}(\mathrm{M})$, where the sgn function is applied element-wise, and define $\mathrm{M}^\sigma_{\mathcal{I}_0}$ to be the sub-matrix of M^σ obtained by considering the rows corresponding to the set $\mathcal{I}$.

Theorem 4.9. *The MLE does not exist if and only if $M^\sigma_{\mathcal{I}_0}$ admits a further row sub-matrix $\mathrm{M}^\sigma_{\mathcal{J}}$ whose nonzero columns each contain elements with opposite signs.*

Proof. Suppose the MLE is not defined. This occurs if and only if there exists $\mathcal{J} \subseteq \mathcal{I}_0$ such that, for each $\mathbf{w} \in \mathcal{L}$, $\mathbf{w}(j) > 0$, with $j \in \mathcal{J}$, implies $\mathbf{w}(j') < 0$ for some other $j' \in \mathcal{J}$, $j \neq j'$. Equivalently, for each positive integer c, the fiber $\mathcal{F}(c\mathbf{n})$ does not contain any element $\mathbf{u}$ such that $\mathbf{u}(j) > 0$, for some $j \in \mathcal{J}$. Because of the connectedness property of the Markov bases (see, for example Sturmfels, 1996, Theorem 5.3), this in turn occurs if and only if no nonzero column of $\mathrm{M}^\sigma_{\mathcal{J}}$ contains only elements having the same signs. ■

If the MLE is not defined, then the vector of sufficient statistics corresponding to $\mathbf{n}$ lies on a face of the marginal cone with facial set $\mathcal{F}$. This does not necessarily imply that there exists any integer point $\mathbf{x}$ in the fiber $\mathcal{Z}(\mathbf{n})$ such that $\mathbf{x}_{\mathcal{F}} > \mathbf{0}$. However, since there are infinitely many rational points in the convex hull of $\mathcal{Z}(\mathbf{n})$ with this property, there exists a big enough positive integer c such that an integer-valued vector $\mathbf{v}$ in $\mathcal{Z}(c\mathbf{n})$ with $\mathbf{v}_{\mathcal{F}} > \mathbf{0}$ can be isolated. These observation is used in the proof the following corollary.

Corollary 4.10. *Let* $\mathrm{M}^{\sigma}_{\mathcal{J}}$ *such that* $|\mathcal{J}|$ *is largest. Then,* $\mathrm{M}^{\sigma}_{\mathcal{J}}$ *is unique and* $\mathcal{F}^c = \mathcal{J}$.

Proof. The uniqueness of $\mathrm{M}^{\sigma}_{\mathcal{J}}$ is clear, for if there were two such sub-matrices, say $\mathrm{M}^{\sigma}_{\mathcal{J}_1}$ and $\mathrm{M}^{\sigma}_{\mathcal{J}_2}$ with maximal $|\mathcal{J}_1| = |\mathcal{J}_2|$, then, letting $\mathcal{J}_3 = \mathcal{J}_1 \cup \mathcal{J}_2$, $\mathrm{M}^{\sigma}_{\mathcal{J}_3}$ would be a sub-matrix whose nonzero columns have each elements of opposite signs and $|\mathcal{J}_3| > |\mathcal{J}_1|$, a contradiction.
Next, by the proof of Theorem 4.9, there is no positive integer c such that $\mathcal{Z}(c\mathbf{n})$ contains a vector $\mathbf{v}$ with $\mathbf{v}_{\mathcal{J}} > \mathbf{0}$. This implies that there is no rational vector $\mathbf{x} \in \mathcal{Z}(\mathbf{n})$ such that $\mathbf{x}_{\mathcal{J}} > \mathbf{0}$. Hence for every point $\mathbf{x} \in \mathrm{convhull}(\mathcal{Z}(\mathbf{n}))$, it must be the case that $\mathbf{x}_{\mathcal{J}} = \mathbf{0}$. Thus $\mathcal{J}^c$ is a facial set and, by maximality of $|\mathcal{J}|$, it must hold that $\mathcal{F}^c = \mathcal{J}$. ■

```
1   : F = I+
2   : do repeat
2.1:    Find u* ∈ B such that sgn(u*_{I0}) > 0 or sgn(u*_{I0}) < 0 and supp(u*_{I0}) is maximal.
2.2:    If no such a vector can be found return F
2.3:    F = F ∪ supp(u*_{I0})
2.4:    if F = I return F
2.5:    I0 = I0 \ supp(u*_{I0})
2.6:    B = B \ {u ∈ B : supp(u_{I0}) ⊆ supp(u*_{I0})}
3   : end
```

Table 2: Pseudo-code for determining the facial set $\mathcal{F}$ associated to an observed table with zero cell counts $\mathcal{I}_0$ and positive cell counts $\mathcal{I}_+$, starting from the Markov basis $\mathcal{B}$.

Combining Theorem 4.9 and Corollary 4.10, an algorithm for the determination of the appropriate facial sets associated to a given observed table can be derived, outlined in Table 2. Furthermore, it can be shown that Markov bases are the minimal subsets of $\mathcal{L}$ for which the algorithm of Table 2 is guaranteed to always provide a correct answer.

We conclude by remarking that both the results above depend on the elements of the Markov basis only through their signs. This in fact reflects the fact that deciding whether the MLE exists and computing the appropriate facial set can be presented as a combinatorial tasks. This viewpoint, which has natural connections with the theory of matroids (see Björner et al., 1999) is stressed in Section 4.5.3.

4.3 Maximizing the Log-Likelihood Function

Except for the case of decomposable models, for which both a closed form expression for the MLE and extended MLE and a very simple efficient algorithm for computing both of them are available (see Section 5.2 and Section 5.2.1), optimization of the log-likelihood function will be performed using the Newton-Raphson algorithm.

It is well known (see, for example, Agresti, 2002) that the Newton-Raphson method is extremely fast and efficient except when MLE fails to exist, a situation in which the procedure becomes unstable. The reason of such undesirable behavior is that, when the MLE is not defined, the "non-estimable" components of the parameter space correspond to the directions of recession of the log-likelihood function which, in turn, are identified by the vectors defining the facial set $\mathcal{F}$

associated with the observed sufficient statistics. As a result, by computing the Newton steps along the steepest directions of increase, the supremum of the log-likelihood is realized in the limit by a sequence of points with norms exploding to infinity. Therefore, as the algorithm progresses, the hessian for the log-likelihood function along the mentioned sequence becomes closer and closer to being singular, making the procedure numerically unstable (see Fienberg and Rinaldo, 2006; Rinaldo, 2006). This problem can be fixed by removing the directions of recession, namely by optimizing a reduced form of the log-likelihood function parametrized by the cells in the facial set $\mathcal{F}$. This is equivalent to identifying the restricted member of the extended exponential family associated to the facial set. Such a restricted log-likelihood function is strictly concave and admits a unique, finite maximizer.

The properties of Newton-Raphson method for computing the MLE of log-linear models are thoroughly described by Haberman (1974, Chapter 3), to which the reader is referred for background. The goal of this section is two-fold and builds on those results. First, efficient algorithms for computing the gradient and hessian of the log-likelihood function under both Poisson and product-multinomial schemes are devised. Secondly, having computed the facial set corresponding to the observed sufficient statistic with any of the methods developed in the previous section, it will be shown how to modify the log-likelihood function in a straightforward way in order to obtain the extended MLE.

4.3.1 Poisson Sampling Scheme

Letting U be the full-column rank design matrix whose column range is the log-linear subspace $\mathcal{M}$, the MLE of the cell mean vector $\mathbf{m}$, with $\boldsymbol{\mu} = \log \mathbf{m} \in \mathcal{M}$, is computed by solving the unconstrained optimization problem

$$\sup_{\mathbf{x} \in \mathbb{R}^k} \ell_{\mathcal{P}}(\mathbf{x}), \tag{50}$$

where $\ell_{\mathcal{P}}(\mathbf{x}) = \mathbf{n}^\top \mathrm{U}\mathbf{x} - \mathbf{1}^\top \exp^{\mathrm{U}\mathbf{x}}$ and the log-likelihood function $\ell_{\mathcal{P}}$ is defined in Equation (2).

By setting $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{U}\mathbf{x}$ and $\mathbf{m}_{\mathbf{x}} = \exp^{\boldsymbol{\mu}_{\mathbf{x}}}$, it can be seen that

$$\begin{aligned} \ell_{\mathcal{P}}(\mathbf{x}) &= \mathbf{n}^\top \boldsymbol{\mu}_{\mathbf{x}} - \mathbf{1}^\top \mathbf{m}_{\mathbf{x}} \\ \nabla \ell_{\mathcal{P}}(\mathbf{x}) &= \mathrm{U}^\top (\mathbf{n} - \mathbf{m}_{\mathbf{x}}) \\ \nabla^2 \ell_{\mathcal{P}}(\mathbf{x}) &= -\mathrm{U}^\top \mathrm{D}_{\mathbf{m}_{\mathbf{x}}} \mathrm{U}, \end{aligned} \tag{51}$$

where $\mathrm{D}_{\mathbf{m}_{\mathbf{x}}}$ is a diagonal matrix whose diagonal elements are $\mathbf{m}_{\mathbf{x}}$. Since $\mathbf{m}_{\mathbf{x}} > \mathbf{0}$ for each $\mathbf{x} \in \mathbb{R}^k$, it is easy to see that the hessian is negative definite on all $\mathbb{R}^k$ which imply that $\nabla \ell_{\mathcal{P}}$ is strictly concave, but not strongly concave, as $\boldsymbol{\mu}(i) \to -\infty$ for any $i \in \mathcal{I}$ implies $\mathbf{m}(i) \to 0$. It is this "weaker" degree of convexity that permits the occurrence of the extended maximum likelihood estimates.

Under the assumption that the MLE is defined, Newton-Raphson method will convergence from any starting approximation $\mathbf{x}_0$ to the unique optimum $\mathbf{x}^*$ for the problem (74). To see this note that the existence and uniqueness of the extended MLE for the restricted exponential family, along with the strict concavity of $\ell_{\mathcal{P}}$, imply that the contour of $\ell_{\mathcal{P}}$ corresponding to the value of $\ell_{\mathcal{P}}(\mathbf{x}_0)$ is a simple closed curve bounding a compact set B. Since the step size algorithm increases the value of $\ell_{\mathcal{P}}$ with each iteration, the sequence of iterations $\{\mathbf{x}_j\}_{j \geq 0}$ all lie inside B. By strong convexity on B, the iterates must converge to a maximum.

At the k-th step of the algorithm, the current approximation $\mathbf{x}_k$, along with Equation (51), is used to compute the Newton direction $\mathbf{d}_k$ by solving the system

$$\nabla^2 \ell_{\mathcal{P}}(\mathbf{x}_k)\mathbf{d}_k = \nabla \ell_{\mathcal{P}}(\mathbf{x}_k).$$

The Cholesky factorization, for example, can be employed to perform such a task.

After the direction has been computed, the stepsize α_k must be determined by either of the line-searching methods described in Section 4.6. To this extent, the function

$$\phi_k(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{d}_k)$$

and possibly its derivatives will have to be repeatedly evaluated. Define $\mathbf{c}_k = \mathrm{U}\mathbf{d}_k$, so that

$$\phi_k(\alpha) = \mathbf{n}^\top \boldsymbol{\mu}_k + \alpha \mathbf{n}^\top \mathbf{c}_k - \mathbf{1}^\top \left(\mathbf{m}_k \cdot \exp^{\alpha \mathbf{c}_k}\right),$$

where $\boldsymbol{\mu}_k = \mathrm{U}\mathbf{x}_k$, $\mathbf{m}_k = \exp^{\boldsymbol{\mu}_k}$ and the dot product operator between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as $\mathbf{z} = (\mathbf{x} \cdot \mathbf{y})$, with $z_i = x_i y_i$. The first and second derivative of ϕ_k are easily computed as

$$\phi_k'(\alpha) = \mathbf{d}_k^\top \left(\mathbf{x} - (\mathbf{m}_k \cdot \exp^{\alpha \mathbf{c}_k})\right) \quad \text{and} \quad \phi_k''(\alpha) = -\sum_i \left(d_i^{(k)}\right)^2 m_i^{(k)} \exp^{\alpha d_i^{(k)}}.$$

After α_k has been evaluated, set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$, so that $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \alpha_k \mathbf{c}_k \in \mathcal{M}$, since $\mathbf{c}_k \in \mathcal{M}$. As a starting point $\mathbf{x}_0$ one can take, for example, $\mathbf{x}_0 = \left(\mathrm{U}^\top \mathrm{U}\right)^{-1} \mathrm{U}^\top \tilde{\boldsymbol{\mu}}$, with $\tilde{\boldsymbol{\mu}} = \log\left(\max(\mathbf{n}, 1)\right)$.

When the MLE does not exist and the extended MLE corresponding to a facial set $\mathcal{F}$ is to be computed, the reduced log-likelihood function is obtained by replacing (50) with the smaller-dimensional optimization problem

$$\sup_{\mathbf{w} \in \mathbb{R}^{k'}} \ell^*_{\mathcal{P}}(\mathbf{w}), \tag{52}$$

where $\ell^*_{\mathcal{P}}(\mathbf{w}) = \mathbf{n}^\top \mathrm{U}^* \mathbf{w} - \mathbf{1}^\top \exp^{\mathrm{U}^* \mathbf{w}}$ and U^* is a $|\mathcal{F}| \times k'$ full-column rank design matrix consisting of any set of linearly independent columns from $\mathrm{U}_{\mathcal{F}}$. To compute U^* from $\mathrm{U}_{\mathcal{F}}$ any of the techniques described in Section 4.4 of the appendix can be used. Note that k' is both the dimension of the natural parameter space for the reduced linear exponential family associated with the log-linear subspace $\mathcal{R}(\mathrm{U}^*)$ and the dimension of the face of the marginal cone containing the observed sufficient statistics in its relative interior. Once the optimum $\mathbf{w}^*$ for (52) is found, the extended MLE is the vector $\widehat{\mathbf{m}}^{\mathrm{e}} \geq \mathbf{0}$ in $\mathbb{R}^{\mathcal{I}}$ with coordinates

$$\widehat{\mathbf{m}}^{\mathrm{e}}(i) = \begin{cases} \exp^{\mathbf{w}^*(i)} & \text{if } \; i \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases}$$

4.3.2 Product-Multinomial Sampling Scheme

When dealing with the product-multinomial sampling scheme, two strategies are available. First, one can take advantage of of the equivalence of the MLE and extended MLE between Poisson and product-multinomial and, provided the sampling subspace $\mathcal{N}$ is contained in the log-linear subspace $\mathcal{M}$, proceed as if Poisson sampling were in fact used. The second possibility is to perform the optimization by parametrizing the log-likelihood function using an appropriate design matrix for $\mathcal{M} \ominus \mathcal{N}$. This second approach is more elaborated because the gradient and hessian of the re-parametrized log-likelihood are harder to obtain, both theoretically and computationally. There are two cases in which the more complicated procedure might be desirable:

- when $\dim(\mathcal{N})$ is very big and a considerable reduction in the dimensionality can be achieved, more than offsetting the computational ease of the Poisson case, despite the increase in complexity needed to obtain the Newton steps;
- when $\mathcal{N} \not\subset \mathcal{M}$, in which case the equivalence of the MLE and extended MLE does not hold.

Recall that, according to Lemma 2.2, it is possible to parametrize the log-likelihood in terms of vectors $\boldsymbol{\beta} \in \mathcal{M} \ominus \mathcal{N}$, as indicated in Equation (6), reported below for convenience,

$$\ell_{\mathcal{L}}(\boldsymbol{\beta}) = (\mathbf{n}, \boldsymbol{\beta}) - \sum_{j=1}^{r} N_j \log(\exp^{\boldsymbol{\beta}}, \boldsymbol{\chi}_j) - \sum_{i \in \mathcal{I}} n_i!.$$

The gradient and hessian are derived as follows. Let $\mathbf{b} = \exp^{\boldsymbol{\beta}}$ and, for $j = 1, \ldots, r$, $\mathbf{b}_j = \{b_i \colon i \in \boldsymbol{\chi}_j\}$ and $N_j = \boldsymbol{\chi}_j^\top \mathbf{n}$. Then,

$$\nabla \ell_{\mathcal{L}}(\boldsymbol{\beta}) = \mathbf{n} - \begin{pmatrix} \left(\frac{N_1}{\boldsymbol{\chi}_1^\top \mathbf{b}}\right) \mathbf{b}_1 \\ \vdots \\ \left(\frac{N_r}{\boldsymbol{\chi}_r^\top \mathbf{b}}\right) \mathbf{b}_r \end{pmatrix} \tag{53}$$

and

$$\nabla^2 \ell_{\mathcal{L}}(\boldsymbol{\beta}) = -\mathrm{diag}\left(\mathrm{H}_1, \ldots, \mathrm{H}_r\right), \tag{54}$$

where, for $j = 1, \ldots, r$,

$$\mathrm{H}_j = \frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}} \left[\mathrm{D}_{\mathbf{b}_j} - \left(\frac{1}{\boldsymbol{\chi}_j^\top \mathbf{b}}\right) \mathbf{b}_j \mathbf{b}_j^\top \right]. \tag{55}$$

The matrix H_j is positive semidefinite with a single null vector $\mathbf{1}$. It follows that $\nabla^2 \ell_{\mathcal{L}}(\boldsymbol{\beta})$ is negative semidefinite with a null space spanned by $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r$, which is just $\mathcal{N}$. Therefore, conclude that $\nabla^2 \ell_{\mathcal{L}}(\boldsymbol{\beta})$ is negative definite on $\mathcal{M} \ominus \mathcal{N}$, for all $\boldsymbol{\beta} \in \mathcal{M} \ominus \mathcal{N}$.

For each $\boldsymbol{\beta} \in \mathcal{M} \ominus \mathcal{N}$, the proof of Lemma 2.2 shows that there exists a corresponding $\boldsymbol{\gamma} \in \mathcal{N}$ such that the vector $\mathbf{c} = \exp^{\boldsymbol{\gamma}}$ satisfies

1. $\mathbf{c}(i) = c_j := \frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}}$, $i \in \boldsymbol{\chi}_j$, $j = 1, \ldots, r$;
2. $\mathbf{b}(i)\mathbf{c}(i) = \mathbf{m}(i)$, $i \in \mathcal{I}$, , with $\mathbf{m}$ being the mean cell vector.

By multiplying and dividing each element of the second vector on the right hand side of (53), it follows that

$$\nabla \ell_{\mathcal{L}}(\boldsymbol{\beta}) = \mathbf{n} - \begin{pmatrix} \frac{1}{c_1}\left(\frac{N_1}{\boldsymbol{\chi}_1^\top \mathbf{b}}\right) c_1 \mathbf{b}_1 \\ \vdots \\ \frac{1}{c_r}\left(\frac{N_r}{\boldsymbol{\chi}_r^\top \mathbf{b}}\right) c_r \mathbf{b}_r \end{pmatrix} = \mathbf{n} - \mathbf{m}.$$

Using a similar trick, equation (55) can be written as

$$\begin{aligned} \mathrm{H}_j &= \frac{1}{c_j}\frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}} c_j \mathrm{D}_{\mathbf{b}_j} - \frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}}\frac{1}{c_j}\frac{1}{c_j(\boldsymbol{\chi}_j^\top \mathbf{b})} c_j \mathbf{b}_j \mathbf{b}_j^\top c_j \\ &= \mathrm{D}_{\mathbf{m}_j} - \frac{1}{N_j}\mathbf{m}_j \mathbf{m}_j^\top, \end{aligned}$$

where $\mathbf{m}_j = \{m_i \colon i \in \boldsymbol{\chi}_j\}$. Using the last display, the expression in (55) becomes

$$\begin{aligned} \nabla^2 \ell_{\mathcal{L}}(\boldsymbol{\beta}) &= -\mathrm{D}_{\mathbf{m}} - \textstyle\sum_{j=1}^{r} \frac{1}{N_j} \mathbf{m}_j \mathbf{m}_j^{\top} \\ &= \mathrm{D}_{\mathbf{m}} \left(\mathrm{I} - \mathrm{P}_{\mathcal{N}}^{\mathbf{m}} \right), \end{aligned} \tag{56}$$

where $\mathrm{P}_{\mathcal{N}}^{\mathbf{m}}$ is the (oblique) orthogonal projection matrix onto $\mathcal{N}$ relative to the inner product $[\cdot, \cdot]$ on $\mathbb{R}^{\mathcal{I}}$ defined by $[\mathbf{x}, \mathbf{y}] = \mathbf{x}^{\top} \mathrm{D}_{\mathbf{m}} \mathbf{y}$. Notice that equation (56) matches Haberman (1974, equation 2.28). For a characterization of maximum likelihood estimation in terms of oblique projections on $\mathcal{M} \ominus \mathcal{N}$, see Haberman (1977). Furthermore, $\nabla^2 \ell_{\mathcal{L}}$ is negative definite on $\mathbb{R}^{k-r}$ but not strongly concave on it. As in the Poisson case, this feature allows for the possibility of a non-existent MLE.

In the case of multinomial sampling scheme with $(\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r)$ being the r orthogonal 0-1 vectors spanning the sampling subspace $\mathcal{N}$ (see Section 2.1.2), it is shown later in Section 4.3.5 that the subspace $\mathcal{M} \ominus \mathcal{N}$ can be generated from a basis U of $\mathcal{M}$ as the column span of the matrix

$$\mathrm{V} = \mathrm{U}_2 - \mathrm{U}_1 \mathrm{W},$$

for a given partition $\mathrm{U} = [\mathrm{U}_1 \;\; \mathrm{U}_2]$ with

$$\mathrm{U}_1 = [\boldsymbol{\chi}_1 \quad \boldsymbol{\chi}_2 \quad \cdots \quad \boldsymbol{\chi}_r]$$

and $\mathrm{W} = \left(\mathrm{U}_1^{\top} \mathrm{U}_1 \right)^{-1} \mathrm{U}_1^{\top} \mathrm{U}_2$. The log-likelihood can be parametrized as

$$\ell_{\mathcal{L}}(\mathbf{x}) = \mathbf{n}^{\top} \mathrm{V} \mathbf{x} - \sum_{j=1}^{r} N_j \log \boldsymbol{\chi}_j^{\top} \exp^{\mathrm{V}\mathbf{x}}, \tag{57}$$

with $\mathbf{x} \in \mathbb{R}^{k-r}$, so that the MLE is obtained by solving the unconstrained optimization problem

$$\sup_{\mathbf{x} \in \mathbb{R}^{k-r}} \ell_{\mathcal{L}}(\mathbf{x}).$$

The considerations of the previous section apply directly to this problem as well, with the principal algorithmic difference being the computation of the function $\ell_{\mathcal{L}}$ and its gradient and hessian, for which formula (76) can be used. Explicitly, for any $\mathbf{x} \in \mathbb{R}^{k-r}$, let $\boldsymbol{\beta}_{\mathbf{x}} = \mathrm{V}\mathbf{x} \in \mathcal{M} \ominus \mathcal{N}$ and, for $j = 1, \ldots, r$, let $\mathbf{b}_{\mathbf{x}}^j = \{\mathbf{b}_{\mathbf{x}}(i) \colon i \in \boldsymbol{\chi}_j\}$ and $N_j = \boldsymbol{\chi}_j^{\top} \mathbf{n}$. Then, using (53) and (76) the gradient is

$$\nabla \ell_{\mathcal{L}}(\mathbf{x}) = \mathrm{V}^{\top} \mathbf{n} - \mathrm{V}^{\top} \begin{pmatrix} \left(\frac{N_1}{\boldsymbol{\chi}_1^{\top} \mathbf{b}_{\mathbf{x}}} \right) \mathbf{b}_{\mathbf{x}}^1 \\ \vdots \\ \left(\frac{N_r}{\boldsymbol{\chi}_r^{\top} \mathbf{b}_{\mathbf{x}}} \right) \mathbf{b}_{\mathbf{x}}^r \end{pmatrix},$$

while, using (76), (54) and (55), the hessian can be expressed as

$$\nabla^2 \ell_{\mathcal{L}}(\mathbf{x}) = -\sum_{j=1}^{r} \mathrm{V}_j^{\top} \mathrm{H}_{\mathbf{x}}^j \mathrm{V}_j, \tag{58}$$

where V_j denotes the rows of V indexed by $\mathrm{supp}(\boldsymbol{\chi}_j)$ and

$$\mathrm{H}_j = \frac{N_j}{\boldsymbol{\chi}_j^{\top} \mathbf{b}_{\mathbf{x}}} \left[\mathrm{D}_{\mathbf{b}_{\mathbf{x}}^j} - \left(\frac{1}{\boldsymbol{\chi}_j^{\top} \mathbf{b}_{\mathbf{x}}} \right) \mathbf{b}_{\mathbf{x}}^j (\mathbf{b}_{\mathbf{x}}^j)^{\top} \right].$$

When the MLE does not exist and the extended MLE for a given facial set $\mathcal{F}$ is sought, the procedure is identical to the Poisson case. Specifically, re-define the restricted log-likelihood function (57) with domain $\mathbb{R}^{k'}$, $k' < k - r$, as

$$\ell^*_{\mathcal{L}}(\mathbf{w}) = \mathbf{n}^\top \mathrm{V}^* \mathbf{w} - \sum_{j=1}^{r} N_j \log \boldsymbol{\chi}_j^\top \exp^{\mathrm{V}^* \mathbf{w}},$$

where V^* is the full-column-rank $|\mathcal{F}| \times k'$ dimensional matrix consisting of any set of linearly independent columns of $\mathrm{V}_{\mathcal{F}}$, isolated using any of the procedures described in Section 4.4. As with the Poisson case, k' is the dimension of both the natural parameter space of the restricted linear exponential family associated to the extended MLE and the dimension of the face of the $\mathrm{cone}(\mathrm{V}^\top)$ identified by the facial set $\mathcal{F}$. Next, solve the smaller-dimensional optimization problem

$$\sup_{\mathbf{w} \in \mathbb{R}^{k'}} \ell^*_{\mathcal{L}}(\mathbf{w}),$$

which is guaranteed to have one finite maximizer $\mathbf{w}^*$. The extended MLE is the vector $\widehat{\mathbf{m}}^{\mathrm{e}} \geq \mathbf{0}$ in $\mathbb{R}^{\mathcal{I}}$ with coordinates

$$\widehat{\mathbf{m}}^{\mathrm{e}}(i) = \begin{cases} \exp^{\mathbf{w}^*(i)} & \text{if } \ i \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases}$$

4.3.3 Efficient Algorithms to Compute $\ell_{\mathcal{L}}$, $\nabla \ell_{\mathcal{L}}$ and $\nabla^2 \ell_{\mathcal{L}}$

In order to compute $\ell_{\mathcal{L}}$, only the term $\boldsymbol{\chi}_j^\top \exp^{\boldsymbol{\beta}_{\mathbf{x}}}$ can be efficiently generated. To this extent, let $\boldsymbol{\kappa}_{\mathbf{x}}$ be the r-dimensional vector whose j-th coordinate is $\boldsymbol{\kappa}_{\mathbf{x}}(j) = \boldsymbol{\chi}_j^\top \exp^{\boldsymbol{\beta}_{\mathbf{x}}}$. Next, note that for a fixed $i \in \mathcal{I}$, at most one of i-th component of the $\boldsymbol{\chi}_j$'s is nonzero. Consequently, while cycling though the components of $\boldsymbol{\beta}_{\mathbf{x}}$, each contributes to exactly one $\boldsymbol{\kappa}_{\mathbf{x}}(j)$. The pseudo-code for the corresponding algorithm is given in Table 14. The same technique can be used to compute $\nabla \ell_{\mathcal{L}}$; since V can be generated row-wise, the only difficulty to address is the generation of the components (53), which is shown in the pseudo-code of Table 15.

The computation of $\nabla^2 \ell_{\mathcal{L}}$ is a bit more elaborate. Re-write equation (58) in the form

$$\nabla^2 \ell_{\mathcal{L}}(\mathbf{x}) = -\sum_{j=1}^{r} \frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}_{\mathbf{x}}} \mathrm{V}_j^\top \mathrm{D}_{\mathbf{b}_{\mathbf{x}}^j} \mathrm{V}_j + \sum_{j=1}^{r} \frac{N_j}{(\boldsymbol{\chi}_j^\top \mathbf{b}_{\mathbf{x}})^2} \mathbf{w}_j \mathbf{w}_j^\top,$$

where $\mathbf{w}_j = \mathrm{V}_j^\top \mathbf{b}_{\mathbf{x}}^j$. The previous display is, in fact, easier to evaluate than formula (58). To efficiently compute $\nabla^2 \ell_{\mathcal{L}}(\mathbf{x})$, the following strategy is proposed: the indices i from the label set $\mathcal{I}$ are generated from the set $\mathrm{supp}(\boldsymbol{\chi}_1)$, then the set $\mathrm{supp}(\boldsymbol{\chi}_2)$, and so on. Following this ordering, $\mathbf{w}_j$ is computed and then $\mathbf{w}_j \mathbf{w}_j^\top$ accumulated before evaluating $\mathbf{w}_{j+1}$. This implies that $\mathbf{w}_{j+1}$ can overwrite $\mathbf{w}_j$, reducing the required storage to a minimum. The details of this algorithm are presented in Table 16

4.3.4 Manipulation and Computations on Design Matrices

Let U be a $n \times p$ design matrix, not necessarily of full rank, for the log-linear models described in Section 3.2 such that $\mathcal{R}(\mathrm{U}) = \mathcal{M}$. Using any of the methods described in Section 3.2 and Tables

8, 9, 10, 11, assume that a procedure `Get_Row` is available that takes as input an index i, which uniquely identifies a cell, and outputs the number of nonzero entries `nz`, the vector coordinates `coord` and vector of values `v` for the corresponding row i of the matrix U, i.e. $\mathbf{u}_i^\top$.

For any p-dimensional vector $\mathbf{x}$, letting $\mathrm{D}_\mathbf{x}$ be the $p \times p$ diagonal matrix whose (i,i)-th entry is x_i, the following operations can be performed efficiently:

- $\mathrm{U}\mathbf{x}$ (see Table 19).
- $\mathrm{U}^\top \mathbf{y} = \sum_i y_i \, \mathbf{u}_i$ (see Table 18).
- $\mathrm{U}^\top \mathrm{D}_\mathbf{x} \mathrm{U} = \sum_i x_i \, \mathbf{u}_i \mathbf{u}_i^\top$ (see Table 17).

The three algorithms above, along with a linear equation solver to be described next, allow to perform a significant amount of computation without retaining U in storage. For example, to compute the oblique projection of a vector $\mathbf{y}$, which is performed repeatedly in the course of the Newton iteration for the computation of the MLE,

$$\mathrm{U} \left(\mathrm{U}^\top \mathrm{D}_\mathbf{x} \mathrm{U} \right)^{-1} \mathrm{U}^\top \mathbf{y},$$

it is possible to proceed as follows:

1. compute $\mathrm{B} = \mathrm{U}^\top \mathrm{D}_\mathbf{x} \mathrm{U}$,
2. compute $\mathbf{w} = \mathrm{U}^\top \mathbf{y}$,
3. solve the system $\mathrm{B}\mathbf{z} = \mathbf{w}$,
4. compute $\mathbf{x} = \mathrm{U}\mathbf{z}$.

Although the above algorithms have been designed to use repeated calls to `Get_Row`, this is by no means necessary, since it might be possible to store the nonzero elements of U in the format produced by `Get_Row`. For example, when the marginal basis from equation (29) is used, each row of U has precisely f nonzero entries, where f is the number of facets of the simplicial complex encoding the corresponding hierarchical model. Since all the nonzero entries are known to be one, only their column indices `coord` must be stored. Thus, each row of U can be represented in $f+1$ locations, one for `nz` and the remaining for `coord`, so that U can be represented with a smaller $n \times (f+1)$ matrix.

4.3.5 A Basis for $\mathcal{M} \ominus \mathcal{N}$ for the Product-Multinomial Case

Consider the product-multinomial sampling scheme case and suppose the r orthogonal vectors $(\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r)$ span the sampling subspace $\mathcal{N}$. Let U_1 be the $|\mathcal{I}| \times r$ matrix whose i-column is $\boldsymbol{\chi}_i$, so that $\mathrm{D} = \mathrm{U}_1^\top \mathrm{U}_1$ is a r-dimensional non-singular diagonal matrix. Let $\mathrm{U} = [\mathrm{U}_1 | \mathrm{U}_2]$ be such that $\mathcal{R}(\mathrm{U}) = \mathcal{M}$. Then

Lemma 4.11. *The columns of the matrix* $\mathrm{V} = \mathrm{U}_2 - \mathrm{U}_1 \mathrm{D}^{-1} \mathrm{U}_1^\top \mathrm{U}_2$ *span* $\mathcal{M} \ominus \mathcal{N}$.

Proof. Orthogonality of $\mathcal{R}(\mathrm{V})$ and $\mathcal{R}(\mathrm{U}_1)$ follows from

$$\begin{aligned} \mathrm{U}_1^\top \mathrm{V} &= \mathrm{U}_1^\top \mathrm{U}_2 - \mathrm{U}_1^\top \mathrm{U}_1 \mathrm{D}^{-1} \mathrm{U}_1^\top \mathrm{U}_2 \\ &= \mathrm{U}_1^\top \mathrm{U}_2 - \mathrm{D}\mathrm{D}^{-1} \mathrm{U}_1^\top \mathrm{U}_2 \\ &= 0. \end{aligned} \tag{59}$$

It only remains to show that $[\mathrm{U}_1|\mathrm{V}]$ span $\mathcal{M}$. Let $\boldsymbol{\mu} = \mathrm{U}_1\mathbf{b}_1 + \mathrm{U}_2\mathbf{b}_2$ be any vector in $\mathcal{M}$. Then,

$$\begin{aligned} \boldsymbol{\mu} &= \mathrm{U}_1\mathbf{b}_1 + (\mathrm{V} + \mathrm{U}_1\mathrm{D}^{-1}\mathrm{U}_1^\top \mathrm{U}_2)\mathbf{b}_2 \\ &= \mathrm{U}_1(\mathbf{b}_1 + \mathrm{D}^{-1}\mathrm{U}_1^\top \mathrm{U}_2\mathbf{b}_2) + \mathrm{V}\mathbf{b}_2, \end{aligned}$$

so that $\boldsymbol{\mu}$ is a linear combination of the columns of U_1 and V. ■

For hierarchical models and product-multinomial sampling scheme with $\mathcal{N} \subset \mathcal{M}$, assume that the design matrix U is such that the first r columns are precisely the vectors $(\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r)$. Let also $\mathrm{B} = \mathrm{D}^{-1}\mathrm{U}_1^\top \mathrm{U}_2$. Then, the i-th row of V is

$$\mathbf{v}_i^\top = (\mathbf{u}_i^{(2)})^\top - (\mathbf{u}_i^{(1)})^\top \mathrm{B},$$

where $(\mathbf{u}_i^{(j)})^\top$ is the i-th row of U_j, $j = 1, 2$. Each row of U_1 contains only one nonzero element, which is 1, therefore, for each i

$$\mathbf{v}_i^\top = (\mathbf{u}_i^{(2)})^\top - \mathbf{b}_l^\top,$$

where $\mathbf{b}_l^\top$ is the l-th row of B and l is the position of the nonzero element in $(\mathbf{u}_i^{(1)})^\top$. Since, by construction, the first r columns of U are precisely U_1, then, for any $\mathbf{i} = (i_1, \ldots, i_K) \in \mathcal{I}$, the corresponding index l is just the first coordinate of the vector `coord` produced by calling the routine `Get_Row` with input argument $(i_1, \ldots, i_K)$. The resulting algorithm is presented in Table 20.

In general, since V is much less sparse than U, the manipulations discussed in the previous section cannot be performed as efficiently. In fact, optimizing over $\mathcal{M} \ominus \mathcal{N}$ rather than the whole $\mathcal{M}$ will be convenient provided that $\dim(\mathcal{M} \ominus \mathcal{N})$ is much smaller than $\dim(\mathcal{M})$.

4.4 Detecting Rank Degeneracy

The present section provides various methods for isolating a set of independent columns from a matrix A. See Stewart (1998) for detailed descriptions and properties of algorithms used below.

4.4.1 Cholesky Decomposition with Pivoting.

For a given squared, positive definite p-dimensional matrix A, the Cholesky decomposition is an upper triangular matrix R with positive diagonal elements, called the Cholesky factor, such that A can be uniquely decomposed like

$$\mathrm{A} = \mathrm{R}^\top \mathrm{R}.$$

The computation of R is simple, numerically stable and can be performed quite efficiently. It encompasses a sequence of p operations such that at the k-th step of the algorithm, the $k \times p$ matrix R_k is obtained, satisfying

$$\mathrm{A} - \mathrm{R}_k^\top \mathrm{R}_k = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{A}_k \end{pmatrix}, \tag{60}$$

where A_k is positive definite of order $p-k$ and $\mathrm{R}_k = \begin{pmatrix} \mathrm{R}_{k-1} \\ \mathbf{r}_k^\top \end{pmatrix}$, so that $\mathrm{R} = \mathrm{R}_p$. The first $(k-1)$ coordinates of the vector $\mathbf{r}_k$ are 0, the k-th coordinate is equal to $r_k = \sqrt{a_{1,1}^{(k-1)}}$ and the last $(p-k-1)$ coordinates are $\frac{a_{1,j}^{(k-1)}}{r_k}$, $j = k+1, \ldots, p$.

A simple modification of the algorithm described above allows to consider matrices that are only positive semidefinite. In fact, it is not necessary to accept diagonal elements as pivots (i.e. as determining the diagonal elements of R). Specifically, suppose that, at the k-th stage of the reduction algorithm represented by equation (60), the pivoting for the next stage is obtained using another diagonal entry of A_k, say $a_{l,l}^{(k)}$, $l \neq 1$, instead of $a_{1,1}^{(k)}$. Let $\mathrm{J}'_{k+1,l}$ be a permutation matrix obtained by exchanging the first and l-th rows of the identity matrix of order $p-k$ so that

$$\mathrm{J}'_{k+1,l} \mathrm{A}_k \mathrm{J}'_{k+1,l}$$

is a symmetric matrix with $a_{l,l}^{(k)}$ in its leading position and set

$$\mathrm{J}_k = \begin{pmatrix} \mathrm{I}_k & 0 \\ 0 & \mathrm{J}'_{k+1,l} \end{pmatrix}.$$

Then, from (60),

$$\mathrm{J}_k \mathrm{A} \mathrm{J}_k - \mathrm{J}_k \mathrm{R}_k^\top \mathrm{R}_k \mathrm{J}_k = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{J}'_{k+1,l} \, \mathrm{A}_k \, \mathrm{J}'_{k+1,l} \end{pmatrix}. \tag{61}$$

The matrix $\mathrm{R}_k \mathrm{J}_k$ differs from R_k only in having its $(k+1)$-th and $(k+l)$-th columns interchanged. Consequently, (61) represent the k-th step of the Cholesky decomposition of $\mathrm{J}_k \mathrm{A} \mathrm{J}_k$ in which $a_{1,1}^{(k)}$ has been replaced by $a_{l,l}^{(k)}$. If interchanges of leading terms are made at each step, with the exception of the last one, the Cholesky factorization will produce an upper triangular matrix R such that

$$\mathrm{J}_{p-1} \mathrm{J}_{p-2} \ldots \mathrm{J}_1 \, \mathrm{A} \, \mathrm{J}_1 \ldots \mathrm{J}_{p-2} \mathrm{J}_{p-1} = \mathrm{R}^\top \mathrm{R}.$$

That is, R is the Cholesky factor of the matrix A with its rows and columns symmetrically permuted according to $\mathrm{J} = \mathrm{J}_{p-1} \mathrm{J}_{p-2} \ldots \mathrm{J}_1$.

If A is positive semidefinite and the algorithm is carried to its k-th stage, it can be shown that A_k is also positive semidefinite. Unless A_k is zero, it will have a positive diagonal element, which may be exchanged into the pivot so that the $(k+1)$ step can be initiated. Among the possible pivoting strategies, one that is particularly suited to problems of rank detection is taking as pivot element the largest diagonal element of A_k, for every stage k of the reduction. This will result into a matrix R such that

$$r_{k,k}^2 = \sum_{i=k}^{j} r_{i,j}^2 \qquad j = k, \ldots, p,$$

so that the diagonal elements of R satisfy $r_{1,1} \geq r_{2,2} \geq \ldots \geq r_{p,p}$. Moreover, if $r_{k+1,k+1} = 0$ for some k, then the Cholesky factor of A will be of the form

$$\mathrm{R} = \begin{pmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} \\ 0 & 0 \end{pmatrix},$$

where R_{11} has order $k = \mathrm{rank}(\mathrm{A})$.

To isolate a set of independent columns form a matrix U the following result can be used.

Proposition 4.12. *Let* J *be a permutation matrix such that* $\mathrm{UJ} = (\mathrm{U}_1, \mathrm{U}_2)$ *with* U_1 *of order* k*. If*

$$(\mathrm{U}_1, \mathrm{U}_2)^\top (\mathrm{U}_1, \mathrm{U}_2) = \begin{pmatrix} \mathrm{A}_1 & \mathrm{A}_2 \\ 0 & 0 \end{pmatrix}^\top \begin{pmatrix} \mathrm{A}_1 & \mathrm{A}_2 \\ 0 & 0 \end{pmatrix},$$

where A_1 *is non-singular of order* k*, then:*

i. *The columns of* U_1 *are linearly independent;*

ii. $\mathrm{U}_2 = \mathrm{U}_1 \mathrm{A}_1^{-1} \mathrm{A}_2$*;*

iii. *the columns of the matrix*

$$\begin{pmatrix} -\mathrm{A}_1^{-1}\mathrm{A}_2 \\ \mathrm{I}_{p-k} \end{pmatrix} \tag{62}$$

form a basis for the null space of UJ*.*

Proof. Since $\mathrm{U}_1^\top \mathrm{U}_1 = \mathrm{A}_1$ is non-singular and positive definite, U_1 has independent columns, proving *i.*. To establish *ii.*, note that

$$\mathrm{rank}(\mathrm{U}_1, \mathrm{U}_2) = \mathrm{rank} \begin{pmatrix} \mathrm{A}_1 & \mathrm{A}_2 \\ 0 & 0 \end{pmatrix} = k,$$

so U_2 can be obtained as a linear combination of columns of U_1, say

$$\mathrm{U}_2 = \mathrm{U}_1 \mathrm{X}. \tag{63}$$

Then, after pre-multiplying both sides by $\mathrm{U}_1^\top$, it follows that

$$\mathrm{X} = (\mathrm{U}_1^\top \mathrm{U}_1)^{-1} \mathrm{U}_1^\top \mathrm{U}_2 = \mathrm{A}_1^{-1} \mathrm{A}_2.$$

To show *iii.*, observed that the matrix (62) has $p - k$ independent columns and, by *i.*, it satisfies

$$\begin{aligned} (\mathrm{U}_1, \mathrm{U}_2) \begin{pmatrix} -\mathrm{A}_1^{-1}\mathrm{A}_2 \\ \mathrm{I} \end{pmatrix} &= -\mathrm{U}_1 \mathrm{A}_1^{-1} \mathrm{A}_2 + \mathrm{U}_2 \\ &= -\mathrm{U}_1 \left(\mathrm{U}_1^\top \mathrm{U}\right)^{-1} \mathrm{U}_1^\top \mathrm{U}_2 + \mathrm{U}_2 \\ &= -\mathrm{U}_2 + \mathrm{U}_2 \\ &= 0, \end{aligned}$$

where the second to the last inequality is justified by (63). Since the null space of $(\mathrm{U}_1, \mathrm{U}_2)$ has dimension $p - k$, the columns of (62) form a basis for it. ■

4.4.2 Gaussian Elimination, Gauss-Jordan Elimination with Full Pivoting and Reduced Row Echelon Form

Proposition 4.12 suggests different ways of dealing with detecting rank degeneracy other than Cholesky decomposition with pivoting. A matrix that as undergone Gaussian elimination is said to be in *row echelon form*. Specifically, a matrix A is in echelon form if:

1. The leading entry in any nonzero row is 1;

2. the leading entry of each nonzero row after the first occurs to the right of the leading entry of the previous row;

3. all zero rows are at the bottom of the matrix.

A matrix A that as undergone Gauss-Jordan elimination is said to be in *reduced row echelon form*. Besides the conditions 1. - 3. above, it satisfies a further requirement:

4. Any column containing a leading term has only one nonzero entry.

Both the procedures require, in order to achieve numerical stability, the permutations of the rows, called partial pivoting.

If the transformed matrix A is not full column rank, it is convenient to consider full-pivoting, which entails finding also permutations matrices $\mathrm{J}_1, \ldots, \mathrm{J}_r$, $1 \leq r < k$ such that, letting $\mathrm{J} = \prod_{i=1}^{r} \mathrm{J}_i$,

$$\mathrm{AJ} = \begin{pmatrix} \mathrm{R} & \mathrm{B} \\ 0 & 0 \end{pmatrix}, \tag{64}$$

where R is upper triangular if A is in row echelon form or is the identity matrix is A is in reduced row echelon form.

Therefore, if the matrix U is not of full-column rank and the echelon form of $\mathrm{U}^\top\mathrm{U}$ satisfies equation (64), then U satisfies the condition of Proposition 4.12. In particular, if reduced echelon form is performed, then a basis for the null space of UJ is given by $\begin{pmatrix} -\mathrm{B} \\ \mathrm{I} \end{pmatrix}$.

The arguments just made can, in principle, be applied directly to the matrix U rather than $\mathrm{U}^\top\mathrm{U}$. In fact, it is easy to see that, if equation (64) holds, with A being the Row Echelon or Reduced Row Echelon of U, then the permuted indices of the columns of R give a subset of the columns of U that are linearly independent. This strategy, however, might not be convenient if the number of rows of U is very big.

4.4.3 LU Factorization

The LU-factorization of a $m \times n$ matrix A, with $m \geq n$, is the representation

$$\mathrm{LU} = \mathrm{PA},$$

where L is a lower triangular (lower trapezoidal if $m > n$) matrix with unit diagonal elements and dimension $m \times n$, U is a $n \times n$ upper triangular matrix and P a $m \times m$ permutation matrix satisfying $\mathrm{P} = \prod_{j=n-1}^{1} \mathrm{P}_i$, with P_j being the permutation matrix that swap the j-th row with the pivot row during the j-the iteration of the the outer loop of the algorithm (Crout's algorithm).

Lemma 4.13. *Let* U *be a* $m \times n$ *matrix, where*$m > n$*. The first* $k < m$ *columns of* U *are linearly independent if and only if* U *admits the following LU factorization, for some permutation matrix* P*:*

$$\mathrm{PU} = \begin{pmatrix} \mathrm{A} & 0 \\ \mathrm{B} & \mathrm{C} \end{pmatrix} \begin{pmatrix} \mathrm{D} & \mathrm{E} \\ 0 & 0 \end{pmatrix}$$

where A *and* D *are lower and upper* $k \times k$ *full-rank matrices, respectively.*

Proof. To prove sufficiency, note that, by hypothesis

$$\mathrm{PU} = \begin{pmatrix} \mathrm{AD} & \mathrm{AE} \\ \mathrm{BD} & \mathrm{BE} \end{pmatrix},$$

so that the first k columns of PU are linearly independent. This clearly is still true for any row permutation and, hence, it is also true for the permutation matrix P^{-1}.

If D and hence A have dimensions different than k, then $\mathrm{rank}(\mathrm{U}) = \mathrm{rank}(\mathrm{PU}) \neq k$, from which necessity follows. ■

Lemma 4.13 can be used as follows. Let $(i_1, \ldots, i_k) \subseteq \{1, \ldots, n\}$ denote the nonzero diagonal entries of D. Then the columns of U labeled by $(i_1, \ldots, i_k)$ are linearly independent and the rank of U is k. Similar arguments can be applied to the matrix $\mathrm{U}^\top \mathrm{U}$, but in this case it is preferable to use the Cholesky decomposition since it is known to be faster and numerically very stable.

4.5 Appendix A: Alternative Methods for Determining Facial Sets

This appendix describes various methods for identifying the facial sets that are alternative to the LP and non-linear optimization procedures we described above.

4.5.1 Maximum Entropy Approach

Identification of the appropriate facial set can be carried out by replacing the linear objective function of the optimization problem (43), with Shannon's entropy function. The new problem is

$$\begin{array}{rccc} & \max -\sum_i y_i \log y_i & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{A}\mathbf{x} \\ & \mathbf{y} & \geq & \mathbf{0} \\ & \mathbf{1}^\top \mathbf{y} & = & 1. \end{array} \tag{65}$$

The strictly concavity of the entropy function and the fact that $\lim_{x \downarrow 0} x \log x = 0$ guarantee that, for the unique maximizer $\mathbf{y}^*$ of 65, $\mathrm{supp}(\mathbf{y}^*)$ is maximal with respect to inclusion. In fact, letting Δ_0 denote the simplex in $\mathbb{R}^{\mathcal{I}_\mathrm{A}}$, the entropy function is maximized over the convex polytope $\mathrm{DC}_\mathrm{A}^{\mathbf{1}} := \mathrm{DC}_\mathrm{A} \cap \Delta_0$. Such intersection is trivial when the MLE exists and is the point $\mathbf{0}$. In this case, the problem is infeasible. Otherwise, due to the strict concavity of the entropy function, the optimum is achieved inside $\mathrm{ri}\left(\mathrm{DC}_\mathrm{A}^{\mathbf{1}}\right)$, which corresponds to the maximal co-face. Note that $\mathrm{DC}_\mathrm{A}^{\mathbf{1}}$ is typically not of full dimension (unless $\mathcal{I}_\mathrm{A} = \mathcal{F}^c$), in which case the maximizer belongs to a relatively open neighborhood inside $\mathrm{DC}_\mathrm{A}^{\mathbf{1}}$.

With $\mathbf{a}_i^\top$ denoting the i-th row of A, the problem (65) can be rewritten in a more compact form by making the constraint $\mathbf{y} = \mathrm{A}\mathbf{x}$ implicit. Then

$$\begin{array}{rccc} & \max H(\mathbf{x}) & & \\ \text{s.t.} & \mathrm{A}\mathbf{x} & \geq & \mathbf{0} \\ & \mathbf{1}^\top \mathrm{A}\mathbf{x} & = & 1, \end{array}$$

where, for $\mathrm{A}\mathbf{x} > \mathbf{0}$, $H(\mathbf{x}) = -\sum_i \mathbf{a}_i^\top \mathbf{x} \log(\mathbf{a}_i^\top \mathbf{x})$, with gradient

$$\nabla H(\mathbf{x}) = -\mathrm{A}^\top \left(\mathbf{1} + \log(\mathrm{A}\mathbf{x})\right)$$

and hessian

$$\nabla^2 H(\mathbf{x}) = -\mathrm{A}^\top \operatorname{diag}(\mathrm{A}\mathbf{x})^{-1}\,\mathrm{A} = -\sum_i \frac{1}{\mathbf{a}_i^\top \mathbf{x}} \mathbf{a}_i \mathbf{a}_i^\top.$$

4.5.2 Maximum Entropy and Newton-Raphson

By taking the log of the negative of the function f of Equation (44), the optimization problem (45) can be represented as an unconstrained geometric program

$$\min \quad \log\left(\textstyle\sum_i \exp^{\mathbf{a}_i^\top \mathbf{x}}\right),$$

which is equivalent to a linearly constrained one,

$$\begin{array}{ll} \min & \log\left(\sum_i \exp(y_i)\right) \\ \text{s.t.} & \mathrm{A}\mathbf{x} = \mathbf{y}, \end{array} \tag{66}$$

with feasible set given by the kernel of the matrix $[\mathrm{I} \quad -\mathrm{A}]$.

If $\mathbf{x}^*$ is the maximizer of the original problem (45), then this is also the minimizer for the geometric program (66), where the infimum can possibly be $-\infty$ (which of course only happens when $\sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x}) = 0$).

The conjugate of the log-sum-exp function appearing in (66) is the negative entropy function restricted to the simplex, given by

$$\begin{cases} \sum_i \nu_i \log \nu_i & \boldsymbol{\nu} \geq 0 \quad \mathbf{1}^\top \boldsymbol{\nu} = 1 \\ \infty & \text{otherwise}, \end{cases}$$

so the dual of the reformulated problem (66) is

$$\begin{array}{ll} \max & -\sum_i \nu_i \log \nu_i \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{\nu} = 1 \\ & \mathrm{A}^\top \boldsymbol{\nu} = \mathbf{0} \\ & \boldsymbol{\nu} \geq 0. \end{array} \tag{67}$$

Proposition 4.14. *If the MLE exists, the problem (67) admits a unique strictly positive solution $\boldsymbol{\nu}^*$. If the MLE does not exist:*

a) if the zeros $\mathcal{I}_\mathrm{A} = \mathcal{F}^c$, then (67) is infeasible;

b) otherwise, the problem (67) is feasible and admits a unique solution $\boldsymbol{\nu}^*$ such that the co-face is given by the coordinates not in $\operatorname{supp}(\boldsymbol{\nu}^*)$.

Proof. Note that, by the properties of the entropy function, any solution $\boldsymbol{\nu}^*$ to the above problem has maximal support among all the non-negative vectors satisfying the equality constraint. Next, by strict concavity of the entropy function, if the problem is feasible, then it admits a unique solution. If the MLE exists, the maximum is attained at a strictly positive point $\boldsymbol{\nu}^* > \mathbf{0}$ by Theorem Stiemke's 4.19.

Suppose instead that the MLE does not exist. If the system $\mathrm{A}\mathbf{x} > \mathbf{0}$ admits a solution, then, by Gordan's Theorem 4.18, there is only one vector $\boldsymbol{\nu}^*$ satisfying the matrix equality constraint:

$\boldsymbol{\nu}^* = \mathbf{0}$. Therefore, in this case the problem is infeasible. This proves *a)*. Otherwise, the solution is given by a vector $\boldsymbol{\nu}^* \gneq \mathbf{0}$. In this case, the coordinates in $\mathrm{supp}(\boldsymbol{\nu}^*)^c$ give the appropriate co-face. In fact, $\mathbf{0} = (\boldsymbol{\nu}^*)^\top \mathrm{A} = (\boldsymbol{\nu}^*)^\top \mathrm{U}_0 \mathrm{X}$ implies that every $\mathbf{d} \in \mathrm{kernel}(\mathrm{U}_+)$ will be orthogonal to a strictly positive convex combination of the rows of U_0 corresponding to the coordinates in $\mathrm{supp}(\boldsymbol{\nu}^*)$. Since the MLE does not exist, there exists a vector $\mathbf{d}_* \in \mathrm{kernel}(\mathrm{U}_+)$ such that $\mathbf{d}_*^\top \mathbf{u}_i = 0$ for all $i \in \mathcal{F}$ and $\mathbf{d}_*^\top \mathbf{u}_i > 0$ for all $i \in \mathcal{F}^c$, that is, $\mathbf{d}_*$ is orthogonal to all strictly positive combinations of rows of U indexed by $\mathcal{F}$. By maximality of $\mathrm{supp}(\boldsymbol{\nu}^*)$, these rows are the ones in U_+ and the ones in $\mathrm{supp}(\boldsymbol{\nu}^*)$. Hence the result in *b)*. ■

4.5.3 Facial Sets and Gale Transform

The same computational methods devised in the previous section for identifying facial sets can be also applied in a different framework, which is essentially based on the idea of the Gale transform, defined below. The duality of this approach will become apparent in the description that follows.

Let B be a matrix whose columns form a basis for $\mathrm{kernel}(\mathrm{U}^\top)$ and B_0 be the sub-matrix obtained from B by considering the rows in $\mathcal{I}_0$. The matrix B is easy to compute, either by transforming A in normal form, or, for the log-linear models described in Section 3, by using Corollary 3.15. It is possible to describe conditions for the existence of the MLE using B rather than U, taking advantage of the notion of the Gale transform. Assume without loss of generality that $\mathbf{1} \in \mathrm{range}(\mathrm{U})$ and let $\mathbf{u}_i^\top$ denote the i-th row of the design matrix U and, likewise, $\mathbf{b}_i^\top$ be the i-th row of the matrix B.

Definition 4.15. The vectors $\{\mathbf{b}_i\}_{i \in \mathcal{I}}$ form a *Gale transform* of the column vector configuration $\{\mathbf{u}_i\}_{i \in \mathcal{I}}$.

Gale transforms (or Gale diagrams) are used to convert combinatorial statements into geometric ones, and vice versa. See Ziegler (1998) and Grünbaum (2003).

Lemma 4.16. *The MLE does not exist if and only if there exists a vector* $\mathbf{y} \gneq \mathbf{0}$ *such that* $\mathrm{B}_0^\top \mathbf{y} = \mathbf{0}$. *Furthermore, for any such a vector* $\mathbf{y}^*$ *with maximal support,* $\mathrm{supp}(\mathbf{y}^*) = \mathcal{F}^c$.

Proof. The MLE does not exist if and only if the sufficient statistic $\mathbf{t}$ belong to a face of the marginal cone generated by $\mathrm{U}^\top$ whose facial set $\mathcal{F}$ satisfies $\mathcal{F}^c \subseteq \mathcal{I}_0$. By Grünbaum (2003, Result 1 on page 88) the set $\mathcal{F}^c$ is co-facial if and only if $\mathbf{0} \in \mathrm{ri}\left(\mathrm{convhull}(\mathrm{B}_{\mathcal{F}^c}^\top)\right)$ which occurs if and only if there exists a $\mathbf{y} > \mathbf{0}$ such that $\mathrm{B}_{\mathcal{F}^c}^\top \mathbf{y} = \mathbf{0}$. Since $\mathcal{F}^c \subseteq \mathcal{I}_0$, the first statement follows. As for the second statement, $\mathcal{F}^c$ must correspond to such a vector $\mathbf{y}$ with maximal support, by the properties of the points in the relative interior of $\mathrm{convhull}(\mathrm{B}_{\mathcal{F}^c}^\top)$. ■

Remark.
In the language of matroids, deciding whether the MLE exists or not is equivalent to the task of deciding whether $\mathrm{B}_0^\top$ is an acyclic vector configuration (see Ziegler, 1998, Chapter 6). See the Remark following Corollary 4.3.

Using Gordan's Theorem 4.18 in conjunction with Lemma 4.16, it can be seen that the existence of the MLE is equivalent to the existence of a strictly positive solution of the linear system of inequalities

$$\mathrm{B}_0 \mathbf{x} \gneq \mathbf{0},$$

which, in fact, is precisely a restatement of Theorem 2.1 in Haberman (1974).

Therefore, arguing as in Section 4.1 (see, in particular, the optimization problem (40)) the appropriate facial set can be identified by solving the non-linear optimization problem

$$\begin{array}{ll} \max & |\mathrm{supp}(\mathrm{B}_0\mathbf{x})| \\ \text{s.t.} & \mathrm{B}_0\mathbf{x} \geq \mathbf{0}. \end{array} \tag{68}$$

If $\mathbf{x}^*$ is any optimal solution of this problem, then $\mathcal{F} = \mathrm{supp}(\mathrm{B}_0\mathbf{x}^*)$. Note the fact that the facial set is given by the points in the support of the optimal solution, unlike the similar problem (41) involving the matrix A.

The geometric representation of problem (68) is analogous to the problem formulated in Equation (41). Consider the polyhedral cone $\mathrm{DC_B} = \{\mathbf{y} \geq \mathbf{0} \colon \mathbf{y} = \mathrm{B}_0\mathbf{x}, \mathbf{x} \in \mathbb{R}^{|\mathcal{I}|-k}\}$. The set $\mathrm{DC_B}$ has a dual interpretation with respect to the $\mathrm{DC_A}$ defined in Section 4.1.1. In fact, the MLE exists if and only if $\mathrm{DC_B}$ contains only strictly positive points. When the MLE does not exist, the face lattice of $\mathrm{DC_B}$ is the opposite of the face lattice of $\mathrm{DC_A}$.

The methods described in the previous sections can, in fact, all be applied to problem (68). The only fundamental difference lies in the interpretations of the optimal solutions, which are given below. The proofs are similar to the cases treated above and are thus omitted.

- **Linear Programming**

$$\begin{array}{lrcl} & \max \mathbf{1}^\top\mathbf{y} & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{B}_0\mathbf{x} \\ & \mathbf{y} & \geq & \mathbf{0} \\ & \mathbf{y} & \leq & \mathbf{1} \end{array} \tag{69}$$

  This is a possibly iterative procedure which is essentially identical to the one used in Section 4.1.1. Replace B_0 with $\mathrm{B}_{\mathrm{supp}(\mathrm{B}_0\mathbf{x}^*)^c}$ at each step until either the objective function is 0 or $\mathrm{supp}(\mathrm{B}_0\mathbf{x}^*)^c = \emptyset$. The former case implies nonexistence of the MLE with $\mathcal{F}^c$ coinciding the row index set of the current B_0, the latter implies existence of the MLE.

- **Maximum Entropy Approach**

$$\begin{array}{lrcl} & \max -\sum_i y_i \log y_i & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{B}_0\mathbf{x} \\ & \mathbf{y} & \geq & \mathbf{0} \\ & \mathbf{1}'\mathbf{y} & = & 1 \end{array} \tag{70}$$

  Nonexistence of the MLE will either produce infeasibility (this occurs if $\mathcal{I}_{\mathrm{B}_0} = \mathcal{F}^c$) or an optimal solution $\mathbf{x}^*$ such that $\mathcal{F}^c = \mathrm{supp}(\mathrm{B}_0\mathbf{x}^*)^c \subsetneq \mathcal{I}_{\mathrm{B}_0}$.

- **Newton-Raphson Procedure**

$$\sup_{\mathbf{x}} f(\mathbf{x}) := -\mathbf{1}^\top \exp(\mathrm{B}_0\mathbf{x})$$

  If the MLE does not exist, then for any sequence $\{\mathbf{x}_n^*\}$ such that $\lim f(\mathbf{x}_n^*) = \sup_{\mathbf{x}} f(\mathbf{x})$, let $\mathbf{y}^* = \lim_n \exp^{\mathrm{B}_0\mathbf{x}_n^*}$. Then, $\mathcal{F}^c = \mathrm{supp}(\mathbf{y}^*)$. When the MLE exists the supremum of the objective function is 0.

- **Maximum Entropy and Newton-Raphson**

$$\begin{array}{ll} \max & -\sum_i \nu_i \log \nu_i \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{\nu} = 1 \\ & \mathrm{B}_0^\top \boldsymbol{\nu} = \mathbf{0} \\ & \boldsymbol{\nu} \geq 0 \end{array} \tag{71}$$

This is directly related to the Gale transform characterization of a facial set. If the MLE exists, the problem is infeasible. When the MLE does not exist, $\mathcal{F}^c = \mathrm{supp}(\boldsymbol{\nu}^*)$.

4.5.4 Matroids and Graver Basis

This section elaborates on the remarks following Corollary 4.3 and Lemma 4.16 and is also directly related to Section 4.2. Conditions for the existence of the MLE and the determination of the facial sets associated to facets of the marginal cone can, in fact, be derived using the language of realizable oriented matroids. Given the reduced practical utility of the methods presented below and the well established nature of the results utilized, exposition will be kept to a minimum. For a complete account see Björner et al. (1999) and Ziegler (1998, Chapter 6).

Let A be a design matrix whose rows span the log-linear subspace and identify it with a collection of $\mathcal{I}$ column vectors. Let $\mathcal{X}_\mathrm{A}$ be the set of minimal (with respect to inclusion) sign vectors of all linear dependencies of A, that is the set of sign vectors of elements of $\mathrm{kernel}(\mathrm{A})$ of minimal support. Then $(\mathcal{I}, \mathcal{X}_\mathrm{A})$ defines an oriented matroid of a vector configurations $\mathcal{I}$ in terms of its signed circuits $\mathcal{X}_\mathrm{A}$. The faces of the corresponding polytope P_A obtained as the convex hull of the columns of A are defined in terms of these circuits. In fact, a set $\mathcal{F}$ defines a face of P_A if and only if, for every signed circuit X, $\mathrm{X}^+ \subset \mathcal{F}$ implies $\mathrm{X}^- \subset \mathcal{F}$. This fact is directly related to Theorem 2.6, part *ii.*, (see also Björner et al., 1999, page 379). Conversely, the signed co-circuit of the matroid $(\mathcal{I}, \mathcal{X}_\mathrm{A})$ form the set of all minimal (with respect to inclusion) sign vectors of value vectors of the matrix A. The combinatorial structure of the polytope P_A can be described in terms of co-circuits as well, since the support of positive co-circuits identifies its co-facets. Circuits and co-circuits are related to each other through the Gale transform of P_A, introduced in Section 4.5.3.

Following Sturmfels (1996), the circuits $\mathcal{X}_\mathrm{A}$ of the toric ideal $\mathcal{I}_\mathrm{A}$ are the irreducible binomials having minimal supports. This definition is equivalent to defining a circuit in $\mathrm{kernel}(A)$ as a non-zero vector whose coordinates are relatively prime and its support is minimal with respect to inclusion (Sturmfels, 1996, page 33). Sturmfels (1996, Proposition 4.11) shows that $\mathcal{X}_\mathrm{A} \subseteq Gr_\mathrm{A}$, where Gr_A denotes the set of all primitive binomials of I_A, called a *Graver basis*.

By combining these notions, it is possible to obtain the co-facets of P_A by computing the signed circuits of the Gale transform of A using the Graver basis of $\mathrm{B}^\top$.

1. Compute the Graver basis of $\mathrm{B}^\top$: Gr_B.
2. Extract the circuits of Gr_B: $\mathcal{X}_\mathrm{B}$
3. Extract the non-negative circuits from $\mathcal{X}_\mathrm{B}$: $\mathcal{X}_\mathrm{B}^+$
4. Compute the sign vectors $\mathrm{sgn}(\mathcal{X}_\mathrm{B}^+)$.
5. The set $\mathrm{sgn}(\mathcal{X}_\mathrm{B}^+)$ gives the co-facets of P_A.

The same steps can be repeated for each facet of $\mathrm{P_A}$.

Remark.
The above characterization, although appealing, has small practical implications, given the current status of available symbolic algebraic algorithms. In fact, the determination Graver bases is a computationally expensive task and Graver bases form even bigger sets than Markov bases, which have shown to become quickly unmanageable as the dimension of the tables increases and, in fact, be arbitrarily complicated (De Loera and Onn, 2006). The methods devised here seem to be, along with the one provided in Section 4.2, more computationally expensive than getting the facet-vertex incidence vectors by using standard polyhedral geometric algorithms, implemented for example in the software `polymake` (Gawrilow, 2000).

Example 4.17. Consider the hierarchical model $\Delta = [12][13][23]$, whose marginal cone has 207 different facets (see Eriksson et al., 2006). The Graver basis of the Gale transform for the corresponding design matrix has 19,197 elements, of which 18,549 are circuits, among which 207 have non-negative entries. The support of those 207 vectors are indeed the facial sets describing the facets of the marginal cone. The computations were performed using `polymake` (Gawrilow, 2000) and `4ti2` (Hemmecke and Hemmecke, 2003). ■

4.6 Appendix B: The Newton-Raphson Method

This section provides a short, specialized description of the Newton-Raphson method which is tailored to the type of applications considered here. Extensive and rigorous treatments of the subject can be found in most books on non-linear optimization. See, for example, Boyd and Vandenberghe (2004).

Let $f\colon \mathbb{R}^k \to \mathbb{R}$ be a real valued function with gradient ∇f and hessian $\nabla^2 f$. The following assumptions will be made:

1. f is strongly concave on any bounded convex subset of $\mathbb{R}^k$;
2. $\nabla^2 f$ is negative definite on all $\mathbb{R}^k$;
3. $\nabla^2 f$ is Lipschitz on $\mathbb{R}^k$.

It can be shown that for the type of applications considered here, all the assumptions hold true.

Newton-Raphson method (or rather the dumped Newton-Raphson method) is a steepest ascent method for finding the maximum of f. It is started with an initial point $\mathbf{x}_0 \in \mathbb{R}^k$ and generates a sequence $\{\mathbf{x}_k\}_{k\geq 0}$ of approximations to the maximum $\mathbf{x}^*$ as follows. Given $\mathbf{x}_k$, let

$$\mathbf{d}_k = -\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

be the *Newton direction* and, for some suitably chosen scalar $\alpha_k > 0$, set

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k.$$

Provided $\mathbf{x}_k \neq \mathbf{x}^*$, the value of the objective function f increases at each iteration, since, letting $\Delta_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and using the fact that $\nabla^2 f$ is negative definite,

$$\nabla f(\mathbf{x}_k)^\top \Delta_k = \alpha_k \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k = -\alpha_k \nabla f(\mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) > 0.$$

By strict concavity, it follows that $f(\mathbf{x}_k) < f(\mathbf{x}_{k+1})$.

Furthermore, if f is strongly concave, there is a maximum $\mathbf{x}^*$ and the sequence α_k can be chosen so that $\lim_k \mathbf{x}_k = \mathbf{x}^*$. Either one of the following algorithms can be used for determining the sequence $\{\alpha_k\}$.

1. *Exact Line Search*. At stage k, choose α_k to maximize the function

$$\phi_k(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$

   Under the assumption of strong concavity, ϕ_k will have a unique maximum which can be found by the one dimensional version of Newton-Raphson method with α as the variable.

2. *Backtracking Line Search*. Choose constants $\sigma > 0$ and $\mu < 1$ and, at stage k, let $i \geq 0$ be the smallest integer for which

$$f(\mathbf{x}_k + \sigma^i \mathbf{d}_k) - f(\mathbf{x}_k) \geq \mu \sigma^i \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k \tag{72}$$

   and set $\alpha_k = \sigma^i$. Condition (72) can always be satisfied, since $\mu < 1$ and, for small α

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) - f(\mathbf{x}_k) \cong \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k.$$

If either one of the above step-size schemes is used, the Newton-Raphson procedure will converge to $\mathbf{x}^*$ for any starting point $\mathbf{x}_0$. Furthermore, for any starting point $\mathbf{x}_0$, there exists a constant K such that, for all $k > K$ $\alpha_k = 1$. This marks the transition from the so-called *dumped Newton phase* to the *quadratically convergent Newton phase*. In fact, for all $k > K$, the iteration occurs at a quadratic rate: there exist a constant M_1 such that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq M_1 \|\mathbf{x}_k - \mathbf{x}^*\|^2.$$

Under the hypotheses that f is strongly convex and $\nabla^2 f$ is Lipschitz, the quadratic nature of convergence assumes the form (see Boyd and Vandenberghe, 2004)

$$|f(\mathbf{x}_k) - f(\mathbf{x}^*)| \leq M_2 \left(\frac{1}{2}\right)^{2^{k-K}},$$

for some constant M_2. Furthermore, the last inequality can be used to compute a bound on the number of iterations in both the dumped and quadratic phase of the Newton-Raphson method.

In the applications considered here, a solution for the constrained optimization problem

$$\begin{array}{ll} \max & g(\boldsymbol{\mu}) \\ \text{s.t.} & \boldsymbol{\mu} \in \mathcal{M} \end{array} \tag{73}$$

is sought, where $\mathcal{M}$ is a linear subspace of dimension k in $\mathbb{R}^{\mathcal{I}}$. If the columns of U form a basis for $\mathcal{M}$, then, letting

$$f(\mathbf{x}) = g(\mathrm{U}\mathbf{x}), \tag{74}$$

the constrained problem (73) becomes the lower-dimensional unconstrained problem

$$\sup_{\mathbf{x} \in \mathbb{R}^k} f(\mathbf{x}). \tag{75}$$

The gradient and hessian of f are easily evaluated from those of g. In fact, for each $\mathbf{x} \in \mathbb{R}^k$, letting $\boldsymbol{\mu}_\mathbf{x} = \mathrm{U}\mathbf{x}$,

$$\nabla f(\mathbf{x}) = \mathrm{U}^\top \nabla g(\boldsymbol{\mu}_\mathbf{x}) \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = \mathrm{U}^\top \nabla^2 g(\boldsymbol{\mu}_\mathbf{x}) \mathrm{U}. \tag{76}$$

Consequently, it is possible to solve (73) by applying the Newton-Raphson method to the simpler problem (75).

4.7 Appendix C: Theorems of Alternatives

The following two theorems of alternatives, whose resemblance with Corollary 4.3 is apparent, were used repeatedly in this section. The proofs of both theorems can be found in Schrijver (1998).

Theorem 4.18 (Gordan's Theorem of Alternatives). *Given a matrix* A*, the following are alternatives:*

1. $\mathrm{A}\mathbf{x} > \mathbf{0}$ *has a solution* $\mathbf{x}$.
2. $\mathrm{A}^\top \mathbf{y} = \mathbf{0}$*,* $\mathbf{y} \gneq \mathbf{0}$*, has a solution* $\mathbf{y}$.

Theorem 4.19 (Stiemke's Theorem of Alternatives). *Given a matrix* A*, the following are alternatives:*

1. $\mathrm{A}\mathbf{x} \gneq \mathbf{0}$ *has a solution* $\mathbf{x}$.
2. $\mathrm{A}^\top \mathbf{y} = \mathbf{0}$*,* $\mathbf{y} > \mathbf{0}$*, has a solution* $\mathbf{y}$.

5 Graph Theory and Extended Maximum Likelihood Estimation

In section 3 we showed how log-linear models, and in particular hierarchical log-linear models, can be defined and described in a combinatorial fashion. Here we further explore the connection between hierarchical log-linear models, graph theory and maximum likelihood estimates and show how graph-theoretical arguments allow in some cases for considerable simplifications, both computational and theoretical. The reader is referred to Lauritzen (1996) for background material.

5.1 Reducible Models

Following Lauritzen (1996), for any hypergraph $\mathcal{H}$, its *reduction*, denoted $\mathrm{red}(\mathcal{H})$ is the antichain consisting of the set of maximal hyperedges of $\mathcal{H}$: $\mathrm{red}(\mathcal{H}) = \{h \in \mathcal{H} \colon h \not\subseteq h', \text{ for any } h' \in \mathcal{H}\}$. The *join* ($\vee$) and *meet* ($\wedge$) of two hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$ are defined respectively as

$$\mathcal{H}_1 \vee \mathcal{H}_2 = \mathrm{red}(\mathcal{H}_1 \cup \mathcal{H}) \quad \text{and} \quad \mathcal{H}_1 \wedge \mathcal{H}_2 = \mathrm{red}\left(\{h_1 \cap h_2, \quad h_1 \in \mathcal{H}_1 \text{ and } h_2 \in \mathcal{H}_2\}\right).$$

The join is said to be direct if $\mathcal{H}_1 \vee \mathcal{H}_2 = \{h\}$, with $h = \left(\bigcup_{h_1 \in \mathcal{H}_1} h_1\right) \bigcap \left(\bigcup_{h_2 \in \mathcal{H}_2} h_2\right)$. Next, consider the partial order $\sqsubseteq$ on the set of hypergraphs on $\mathcal{K}$ implied by the inclusion order of the hyperedges. Specifically, given two hyper-graphs $\mathcal{H}$ and $\mathcal{A}$, $\mathcal{A} \sqsubseteq \mathcal{H}$ when, for every $a \in \mathcal{A}$, $a \subseteq h$ for some $h \in \mathcal{H}$.

Consider the abstract simplicial complex Δ with base set $\mathcal{K} = \{1, \ldots, K\}$ from Section 3 and recall the convention of identifying Δ with its chain of maximal hyper-edges or facets.

Definition 5.1. An abstract simplicial complex Δ is *reducible* when there exist two abstract simplicial complexes Δ_1 and Δ_2 and a face of s of Δ such that

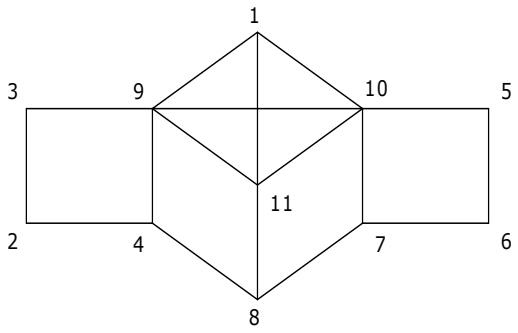
1. $\Delta_1, \Delta_2 \sqsubseteq \Delta$;
2. $\Delta_1 \vee \Delta_2 = \Delta$;
3. $s = \bigcup_{d \in \Delta_1} \cap \bigcup_{d \in \Delta_2}$.

Then, Δ is said to admit a *decomposition* (Δ_1, s, Δ_2) with separator s.

Alternatively, Δ admits a decomposition into the hypergraphs Δ_1 and Δ_2 with separator s if $\Delta = \Delta_1 \vee \Delta_2$ and $\Delta_1 \wedge \Delta_2 = s$, where $d_1 \cap d_2 \subseteq s$ for all $d_1 \in \Delta_1$, $d_2 \in \Delta_2$ and $d_1^* \cap d_2^* = s$ for some $d_1^* \in \Delta_1$, $d_2^* \in \Delta_2$ (see Lauritzen, 2002).

The process can be carried out recursively on each of the two sub-complexes Δ_1 and Δ_2 and, at termination, will produce a decomposition of Δ into sub-complexes and separators, which, without loss of generality, can be ordered in a natural way and labeled as $\Delta_1, \ldots, \Delta_p$ and $s_2, \ldots, s_p$, respectively (notice that, by construction, the number of separators, counting possible multiplicities, is one less than the number of separating sub-complexes).

Figure 4: Example of a reducible graph, from Tarjan (1985, Figure 3).



Example 5.2 (Reducible models). The model $\Delta = [145][12][23][34]$ is graphical, non-decomposable and reducible, with separating sub-complexes $\Delta_1 = [145]$ and $\Delta_1 = [12][23][34][14]$. The model $\Delta = [124][23][13]$ is non-graphical (hence non-decomposable) and reducible, with separating sub-complexes $\Delta_1 = [124]$ and $\Delta_1 = [12][23][13]$. Figure 4 shows a reducible graphical model on 11 vertices corresponding to the hierarchical model model

$$\Delta = [23][24][49][39][19\ \ 10\ \ 11][48][8\ \ 11][78][7\ \ 10][67][56][5\ \ 10].$$

with separators $\{4, 9\}$, $\{7, 10\}$ and $\{9, 10, 11\}$ (see Tarjan, 1985). ■

The reducibility properties of abstract simplexes representing hierarchical log-linear models translate into important factorization properties for the cell mean vector function and the extended MLE, which we describe below. For a hierarchical log-linear model Δ, we denote with $\mathbf{m}^{\Delta}(i) =$

$\mathbf{m}(i)$ the i-th component of the corresponding cell mean vector. Accordingly, if $\mathcal{A} \sqsubseteq \Delta$ is an abstract simplicial complex for which $A = \bigcup_{a\in\mathcal{A}}$ is allowed to be a strict subset of $\mathcal{K}$, we denote with $\mathbf{m}^{\mathcal{A}}(i_A)$ the i_A-th component of cell mean vector corresponding to the log-linear model $\mathcal{A}$, where $i_A \in \bigotimes_{k\in A} I_k$. By Lemma 5.6 in Haberman (1974), for any $i \in \mathcal{I}$, the relation between $\mathbf{m}(i)$ and $\mathbf{m}^{\mathcal{A}}(i_A)$ is given by the identity

$$\mathbf{m}(i) = \frac{\mathbf{m}^{\mathcal{A}}(i_A)}{\prod_{k\notin A} I_k}.$$

The notation is consistent with the case $\mathcal{A} = \{a\}$, with $a \subset \mathcal{K}$, for which Equation (8), implies that $\widehat{\mathbf{m}}^a(i_a) = \mathbf{n}(i_a)$ where $\widehat{\mathbf{m}}^a$ is the MLE for the hierarchical model with $a \subset \mathcal{K}$ as the unique generating class. Note that, if $s = \emptyset$, this notation gives $\mathbf{m}^s(i_s) = \sum_i \mathbf{m}(i)$. The main result of this section is the following theorem which gives explicit factorization formula for the cell mean vector of a reducible model in terms of the cell mean vectors for the decomposing sub-models.

Theorem 5.3. *Let Δ be reducible with decomposing sub-complexes $\Delta_1, \ldots, \Delta_p$ having underlying base sets $D_j = \bigcup_{d\in\Delta_j} d$, $j = 1, \ldots, p$ and separators $s_2, \ldots, s_p$. Then, the toric variety Δ consists of the point-wise limit closure of points $\mathbf{m}$ that factorize as*

$$\mathbf{m}(i) = \frac{\prod_{j=1}^{p} \mathbf{m}^{\Delta_j}(i_{D_j})}{\prod_{j=2}^{p} \mathbf{m}^{s_j}(i_{s_j})} \qquad i \in \mathcal{I}, \tag{77}$$

where the convention $\frac{0}{0} = 0$ is used for the above ratios.

Proof. For a strictly positive cell mean vector $\mathbf{m}^{\Delta}$, Equation (77) follows by applying recursively (Haberman, 1974, Lemma 5.8) and noting that the decomposing sub-complexes Δ_1 and Δ_2 of a reducible complex Δ can always be chosen to be disjoint, so that Δ is the direct join of Δ_1 and Δ_2.

It remains to be shown that any vector in the point-wise limit closure also satisfy (77). Let $L(\Delta_j)$ denote the face lattice (i.e. the set of all facial sets) of the marginal cone for the sub-model Δ_j, $j = 1, \ldots, p$, and $L(\Delta)$ the face lattice for Δ. Since, for any decomposing sub-model Δ_j, $j = 1, \ldots, p$, it is possible that D_j is strictly smaller than $\mathcal{K}$, let $\mathcal{I}_{D_j} = \bigotimes_{k\in D_j} \mathcal{I}_k$ and define the map $\tau_j : 2^{\mathcal{I}_{D_j}} \to 2^{\mathcal{I}}$ given by

$$\tau_j(\mathcal{D}) = \left\{ i \in \mathcal{I} \colon i_{D_j} \in \mathcal{D} \right\},$$

where $\mathcal{D} \subseteq \mathcal{I}_{D_j}$, so that $\tau_j(\mathcal{D}) = \mathcal{I}$ if $D_j = \mathcal{K}$. Equation (77) says that if $\mathcal{F} \in L(\Delta)$, then there exist sets $\mathcal{F}_j \subseteq \mathcal{I}_{\mathcal{D}_j}$, such that $\mathcal{F}_j \in L(\Delta_j)$ and

$$\mathcal{F} = \bigcap_{j=1}^{p} \tau_j(\mathcal{F}_j). \tag{78}$$

By Theorem thm:toricsummary, part *ii.*, all the points in the toric variety associated with the hierarchical model Δ are in one-to-one correspondence with the points in the marginal cone of Δ. Furthermore, the supports of the points of the variety coincide with the facial sets of the marginal cone, hence satisfying Equation (78). Since any point $\mathbf{m}$ on boundary of the associated variety is the point-wise limit of positive mean vectors factorizing as (77) and, at the same time, its support satisfies Equation (78), it must then be the case that $\operatorname{supp}(\mathbf{m}) = \bigcap_{j=1}^{p} \tau_j \left(\operatorname{supp}(\mathbf{m}^{\Delta_j}) \right)$. ■

An immediate consequence of Theorem 5.3 is that, in order to the extended MLE $\widehat{\mathbf{m}}$, it is sufficient to compute the extended MLE of the log-linear models implied by the decomposing sub-complexes and then "glue" them together according to the factorization formula

$$\widehat{\mathbf{m}}(i) = \frac{\prod_{j=1}^{p} \widehat{\mathbf{m}}^{\Delta_j}(i_{D_j})}{\prod_{j=2}^{p} \mathbf{n}(i_{s_j})} \qquad i \in \mathcal{I}, \tag{79}$$

Notice that $\mathbf{n}(i_{s_j}) = 0$ for some separator s_j implies that there exists a set $D_j \supset s_j$ such that $\mathbf{n}(i_{D_j}) = 0$, so that the numerator of (79) is 0 as well and, because of our convention, the resulting value of $\widehat{\mathbf{m}}(i)$ is also null. Although not strong enough to guarantee a closed form representation of the extended MLE, the factorization property (79) have significant practical implications. In fact, it allows for the possibility of breaking down the task of computing the extended MLE for Δ into simpler and computationally less demanding distinct tasks of computing the extended MLE for simpler sub-models.

Remarks.

1. Equation (77) also appears in Dobra and Fienberg (2000) for the subclass of graphical reducible models.

2. When the MLE exists, Equation (79) follows by recursive application of Haberman (1974, Lemma 5.9) or Lauritzen (1996, Proposition 4.14).

3. The previous result shows that the combinatorial structure of the marginal cone for a reducible hierarchical log-linear model is simultaneously determined by the simpler combinatorial structures of the log-linear sub-models. In particular, the points of the toric variety associated to a reducible Δ are obtained by "gluing" (Fulton, 1978; Oda, 1988) together the points in the varieties associated to the sub-models Δ_j, $j = 1, \ldots, p$.

4. The proof of Theorem 5.3 highlights the fact that the face lattice of the marginal cone can be broken down into less complex face lattices corresponding to simpler marginal cones, one for each sub-model; then, the facial sets associated to the original model can be recovered from the conjunction of the facial sets of these smaller pieces. In virtue of the homeomorphism between the toric variety and the marginal cone (Theorem 2.6, part *ii.*) the multiplicative factorization formula (77), satisfied by all point in the hyper-surface defined by the toric variety, has an equivalent polyhedral representation. This correspondence is the linear map given by the Minkowski addition of the marginal cones determined by the decomposing sub-complexes, embedded into the ambient space of the marginal cone for Δ. See Lemma 8 in Eriksson et al. (2006) for details.

Example 5.4. Bishop et al. (1975, Section 3.7.3: *Fitting by Stages*) observed that if there are independent variables, i.e. variables appearing alone in only one generating class, the cell mean vector factorizes in such a way that the maximum likelihood computations can be performed in a stepwise fashion. One of the examples they produced is the model $\Delta = [123][234][134][5]$, whose corresponding MLE is

$$\widehat{m}_{ijklm} = \widehat{m}_{ijkl+} \frac{n_{++++m}}{N}, \tag{80}$$

where $N = n_{+++++}$ and the authors' notation for summing over factors is used. For this model, the MLE can be computed in two stages. First, the data are collapsed summing over the variable

labeled as "5" and then a model is fitted to the condensed table. The resulting fitted values are then adjusted proportionally by the margins of the variable left out in the first stage. This (non-graphical) model is in fact reducible with the empty set as the separator for the two decomposing sub-complexes $\Delta_1 = [123][234][134]$ and $\Delta_2 = [5]$. The MLE formula (80) found by the authors is precisely Equation (79). ■

5.1.1 Decomposing Simplicial Complexes

In order to utilize Theorem 5.3 in practice, algorithms to test the reducibility of Δ and to perform its decomposition are needed. The degree of efficiency of available procedures depend on Δ being graphical or not. Recall from Section 3 that, for a hierarchical log-linear model Δ, $\mathcal{G}(\Delta)$ denotes the corresponding interaction graph and Δ is graphical if the cliques of $\mathcal{G}(\Delta)$ are the facets of Δ.

If Δ is graphical, there exist algorithms capable of providing optimal decompositions of the underlying interaction graph, as described in Tarjan (1985) and Leimer (1993). See Dobra and Fienberg (2000) for a review of these methods and for applications to Fréchet bounds. The procedure described in Dobra and Fienberg (2000, pages 11888–11889) will produce a perfect sequence of subsets of the K factors. These subsets will, in turn, each determine subgraphs of $\mathcal{G}(\Delta)$ which are the interaction graphs of the (necessarily graphical) sub-model Δ_j's forming the partitioning set of Δ. The optimality of these algorithms guarantees that the decomposition is both unique and as thorough as possible, or, in the language of graph theory, that, for each $1 \leq j \leq p$, $\mathcal{G}(\Delta_j)$ is a *maximal prime* subgraph of $\mathcal{G}(\Delta)$ (see Leimer, 1993).

When Δ is not graphical, designing a reduction procedure is more problematic and, to our knowledge, none of the available procedures known to the writer have been demonstrated to produce a solution that is unique or even optimal. Geng (1989) proposed a simple decomposition algorithm for hypergraphs which extends Graham's procedure for decomposable hypergraphs. Geng's algorithm visits each facet of Δ and tests whether it contains a face that is a separator for a decomposition of Δ. As soon as one is found, Δ is decomposed accordingly and the procedure is repeated recursively on the two sub-complexes. Badsberg (1995) instead elaborated on the results of Tarjan (1985) and Leimer (1993) for optimally decomposing the interaction graph $\mathcal{G}(\Delta)$. The extension is based on the observation that if Δ is decomposable with respect to one of its faces s, then s must be a separator of $\mathcal{G}(\Delta)$ as well (the opposite is not true, as separators of $\mathcal{G}(\Delta)$ can in fact be bigger than any facet of a non-conformal Δ). Badsberg (1995) showed that, once the optimal decomposition of the interaction graph $\mathcal{G}(\Delta)$ into prime subgraphs is available, then the minimal separators of $\mathcal{G}(\Delta)$ resulting from such a decomposition which are also faces of Δ are the separators of a maximal decomposition of Δ, where maximality is meant to indicate that the resulting sub-complexes cannot be decomposed any further.

The program `CoCo` (Badsberg, 1995) provides a different implementation of the IPF algorithm and exploits the properties of decomposable and reducible models.

5.2 Decomposable Models

Decomposable log-linear models form a very well-behaved class of graphical models parametrizing families of distributions which factorize according to independence or conditional independence statements. See Lauritzen (1996, Chapter 3) for a thorough description of the Markov properties on decomposable models and Geiger et al. (2006) for their algebraic statistics generalization. For decomposable models, the cell mean vector can be expressed as a rational function of its margins.

Remarkably, the both the MLE and the extended MLE retain such a closed form representation. As a result, the determination of the existence of the MLE and the computation of the extended MLE is almost immediate. The computational consequences of these results are exploited in Section 5.2.1.

The closed form representation is the analytic representation of a very strong graph-theoretic property encoded in the facets of Δ, namely decomposability. A hypergraph is decomposable if it either consists of one hyperedge or can be obtained by direct joins of hypergraphs with fewer hyperedges. Alternatively, a hypergraph is decomposable if its hyperedges are the cliques of a decomposable graph.

Definition 5.5. A hierarchical log-linear model Δ is decomposable if the facets of Δ form a decomposable hypergraph.

Decomposable simplicial complexes are special cases of reducible simplicial complexes in which all the separating sub-complexes are simplices (i.e. they consist each of one class), which must then correspond to the facets of Δ. Consequently Theorem 5.3 applies and produces specialized version of Equations (77) and (79). For completeness, we state it as a separate result.

Corollary 5.6. *The toric variety corresponding to a decomposable model Δ consists of the point-wise limit closure of points $\mathbf{m}$ that factorize as*

$$\mathbf{m}(i) = \frac{\prod_{d\in\Delta} \mathbf{m}^d(i_d)}{\prod_{s\in\mathcal{S}} \mathbf{m}^s(i_s)^{\nu(s)}} \tag{81}$$

and the extended MLE of $\mathbf{m}$ is

$$\widehat{\mathbf{m}}(i) = \frac{\prod_{d\in\Delta} \mathbf{n}(i_d)}{\prod_{s\in\mathcal{S}} \mathbf{n}(i_s)^{\nu(s)}}, \tag{82}$$

where $\frac{0}{0} = 0$, $\mathcal{S}$ is the set of separators and $\nu(s)$ is the multiplicity of $s \in \mathcal{S}$ in any perfect ordering of the set of cliques Δ.

See Section 5.2.2 for a definition and determination of perfect ordering.

When the MLE exists, Equation (82) is a renown important result, proved using different arguments by Haberman (1974, Theorem 5.1) and Lauritzen (1996, Proposition 4.18), the former only under the assumption that the MLE exists.

Haberman (1974) called the numbers $\nu(s)$, $s \in \mathcal{S}$, the *adjusted replication numbers*. Formulas (81) and (82) are equivalent to

$$\mathbf{m} = \frac{\prod_{d\in\Delta} \mathbf{m}^d(i_d)}{\prod_{f\in\mathcal{F}(\Delta)} \mathbf{m}^f(i_f)^{\nu(f)}} \tag{83}$$

and

$$\widehat{\mathbf{m}} = \frac{\prod_{d\in\Delta} \mathbf{n}(i_d)}{\prod_{f\in\mathcal{F}(\Delta)} \mathbf{n}(i_f)^{\nu(f)}}, \tag{84}$$

respectively, where $\mathcal{F}(\Delta) = \{d \cap d' \colon d, d' \in \Delta, d \neq d'\}$ is the *intersection class* associated to the hyper-graph Δ.

An important consequence of the factorization property of the cell mean vector and its MLE for a decomposable model is the fact that MLE exists if and only if the table margins are positive. Let

$$\mathrm{W}_\Delta = \bigoplus_{d\in\Delta} \mathrm{W}_d, \tag{85}$$

where the marginal bases matrices W_h, $h \in 2^{\mathcal{K}}$, are defined in Section 3.2 and the sets d indicate the generating classes, i.e. the facets of Δ.

Corollary 5.7. *For a decomposable log-linear model Δ, the MLE exists if and only if the margins are all positive: $\mathbf{n}(i_d) > 0$, for each $d \in \Delta$ and $i \in \mathcal{I}$. Equivalently, the MLE exists if and only if $\mathrm{W}_\Delta^\top \mathbf{n} > \mathbf{0}$.*

From the geometrical viewpoint, this is equivalent to a small combinatorial structure of the marginal cone associated to the design matrix W_Δ, for any decomposable Δ.

Corollary 5.8. *Let W_Δ be the design matrix (85) associated to the decomposable log-linear model Δ. The facial set for the facets of the corresponding marginal cone are the support sets of the columns of W_Δ.*

Proof. There is more than one way of proving the statement. One strategy is to notice that, by the definition of facial set given in Section 2.2, the supports of the rows of $\mathrm{W}_\Delta^\top$ form facial sets for the facets of marginal cone. Then, the result will follow once it is shown that those are the only facets. This can be accomplished with the same arguments used in the proof of Theorem 5.3. Specifically, by (78) it is sufficient to look at the facial sets for the decomposing sub-complexes. Since each of these sub-complexes is a facet of Δ, Lemma 8 in Eriksson et al. (2006) implies that we only need to study the facial sets for the marginal cones generated by the matrices $\mathrm{W}_d^\top$, for all facets d of Δ. Since, for each $d \in \Delta$, the columns of W_d contain only one non-zero entry, then, using again the definition of facial sets, we conclude that the facial sets for the facets of $\mathrm{cone}(\mathrm{W}_d^\top)$ are precisely the columns of W_d. By (78) and (85), the claim follows.

Alternatively, combine Corollary 3.20 of Lauritzen (1996) and Lemma 1 of Geiger et al. (2006) and note that *nice* sets for the design matrices $\mathrm{W}_\Delta^\top$ obtained as in Equation (85) are the complements of the supports of all columns of W_Δ. ■

Note that similar conclusions can also be derived from Theorem 7.2 in Haberman (1974), a more general and rather involved result which applies to incomplete separable tables and decomposable log-linear models.

5.2.1 The Iterative Proportional Fitting Algorithm

The iterative proportional fitting (IPF) algorithm (known also as iterative proportional scaling) is a widespread method for computing the MLE by performing cyclic iterative partial maximization of the likelihood. It was originally proposed by Deming and Stephan (1940) and further developed by various authors in the context of maximum likelihood estimation for discrete distributions and contingency tables. See Bishop et al. (1975) and Lauritzen (1996) for a description of the procedure and its properties. The IPF procedure has also a natural information-theoretic interpretation: Darroch and Ratcliff (1972) and Csiszár (1975), and more recently, Ruschendorf (1995).

IPF is widely utilized in practice mainly because of two very appealing features:

1. It is very simple to implement and requires minimal memory storage and computational effort. See, for example, the Fortran version of the IPF algorithm written by Haberman (1972, 1976), which is the implemented in the current `R` routine `loglin`.

2. It is guaranteed to eventually reach any arbitrary small neighborhood of the optimum, even when the MLE is not defined, as proved in (Lauritzen, 1996, Theorem 4.13).

Let $\Delta = \{d_1, \ldots, d_f\}$ be a hierarchical log-linear model with facets $d_j \subset \mathcal{K}$, each defining a marginal configuration. Starting from any positive cell mean vector $\mathbf{m}^{(0)}$ such that $\log \mathbf{m}^{(0)} \in \mathcal{M}_\Delta$, the IPF algorithm produces a sequence of points $\{\mathbf{m}^{(j)}\}_{j\geq 0}$ satisfying $\log \mathbf{m}^{(j)} \in \mathcal{M}_\Delta$ for each $j \geq 0$ and $\lim_j \mathbf{m}^{(j)} = \widehat{\mathbf{m}}$. Each point $\mathbf{m}^{(j)}$, with $j > 0$, is obtained by performing a cycle of f sequential adjustments of the margins of $\mathbf{m}^{(j-1)}$ which are proportional to the observed margins. The pseudo-code for the IPF algorithm is given in Table 3.

```
1       : m̂ = 1_I
2       : do repeat
2.1     :     m^0 = m̂
2.2     :     for j = 1 to f
2.2.1   :         m^(j)(i) = m^(j-1)(i) n(i_{d_j}) / m^(j-1)(i_{d_j})
2.3     :     end
2.4     :     m̂ = m^(f)
2.5     :     if distance(W_Δ m̂ − W_Δ n) < tol then return m̂
3       : end
```

Table 3: Pseudo-code for the IPF algorithm. It requires the design matrix W_Δ given in (85), the specification of the function `distance` to measure the discrepancy between the observed and fitted margins and of the maximum deviation allowed `tol`. The convention $\frac{0}{0} = 0$ is used.

The main drawback of the IPF is the fact that it usually exhibits a slow rate of convergence (see, for example Agresti, 2002), typically much slower than procedures based on Newton-Raphson method. This is observed to be particularly true when the MLE is not defined as illustrated in the examples of Fienberg and Rinaldo (2006) and (Rinaldo, 2005, Chapter 1), although no precise results are available. However, there is at least one notable exception for which IPF is to be preferred: the case of a decomposable log-linear model.

As indicated in the previous section, for decomposable models, the MLE and extended MLE have closed form and, in particular, can be expressed as a rational function of the margins. These properties are very important from the practical and computational point of view, since a straightforward modification of equation (82) or its equivalent version (84) will produce the MLE or extended MLE in just one cycle of the IPF algorithm. For a subset $a \subset \mathcal{K}$ of the factors (typically a facet or a separator of Δ), let

$$\mathcal{N}_a = \left\{ i_a \in \bigotimes_{k\in a} I_k \colon \mathbf{n}(i_a) = 0 \right\}$$

be the set holding the coordinates of null observed a-margins. Then, the extended MLE associated to an observed table $\mathbf{n}$ and a decomposable model Δ is

$$\widehat{\mathbf{m}}(i) = \begin{cases} 0 & \text{if} \quad \exists d \colon i_d \in \mathcal{N}_d \\ \dfrac{\prod_{c\in\Delta} \mathbf{n}(i_c)}{\prod_{s\in\mathcal{S}} \mathbf{n}(i_s)^{\nu(s)}} & \text{otherwise} \end{cases} \tag{86}$$

or, equivalently,

$$\widehat{\mathbf{m}}(i) = \begin{cases} 0 & \text{if} \quad \exists d\colon i_d \in \mathcal{N}_d \\ \frac{\prod_{d\in\Delta} \mathbf{n}(i_d)}{\prod_{f\in\mathcal{F}(\Delta)} \mathbf{n}(i_f)^{\nu(f)}} & \text{otherwise.} \end{cases} \tag{87}$$

Furthermore, as remarked by Lauritzen (1996, Proposition 4.35), the determination of the dimension of the log-linear model sub-space $\mathcal{M}_\Delta$ corresponding to a decomposable model can be expressed in closed form, even for a restricted sub-model associated to an extended MLE:

$$\dim(\mathcal{M}_\Delta) = \sum_{d\in\Delta} \left(\prod_{k\in d} I_k - |\mathcal{N}_d| \right) - \sum_{s\in\mathcal{S}} \nu(s) \left(\prod_{k\in s} I_k - |\mathcal{N}_s| \right).$$

Although of immediate applicability, Equations (86) and (87) require the knowledge of the sets of separators or the adjusted replication numbers. Fortunately, this is by no means necessary. In fact, there always exists an ordering, called *perfect*, of the facets of any decomposable model Δ such that one cycle of the IPF algorithm of Table 3 will return both (86) and (87) in just one iteration, provided the marginal updates are performed according to such an ordering.

5.2.2 IPF and Perfect Orderings

Given a graphical hyper-graph $\mathcal{H}$ on K nodes, a numbering $(v_1, \ldots, v_K)$ of the nodes is a *perfect numbering* if the sets $C_1 = \{v_1\}$ and, for $j = 2, \ldots, K$, $C_j = \mathrm{cl}(v_j) \cap (v_1, \ldots, v_{j-1})$, all induce complete sub-graphs of the interaction graph of $\mathcal{H}$. The associated sequence $(C_1, \ldots, C_K)$ of sets is called *perfect* since all the C_j's are complete and their ordering satisfies the *running intersection property*: for all $1 < j \leq K$ there exists a $i < j$ such that $C_j \cap (C_1 \cup \ldots \cup C_{j-1}) \subseteq C_i$. The sets $S_j = C_j \cap (C_1 \cup \ldots \cup C_{j-1})$ are called the separators of the sequence.

Even if neither the perfect numbering of the nodes nor the corresponding perfect sequence of cliques is unique, they are defining features of decomposable graphs and hypergraphs. In fact, it is well known (see Lauritzen, 1996) that a graph is decomposable if and only if its vertices have a perfect numbering, if and only if the hypergraphs of its cliques can be numbered to form a perfect sequence.

The gain in efficiency obtained by combining the IPF algorithm with perfect sequences of margins was prove by Haberman (1974, Theorem 5.3). That result, along with the consideration about the extended MLE for decomposable models is summarized in the next theorem.

Theorem 5.9. *For a decomposable log-linear model with its cliques ordered according to a perfect sequence, the IPF algorithm from Table 3, using $\mathbf{1}_\mathcal{I}$ as a starting value, returns both the MLE and extended MLE in one cycle of iterations.*

5.2.3 Deciding the Decomposability of a Hypergraph

In order to exploit the computational ease that comes with decomposable models it must be verified first that the simplicial complex Δ is in fact associated with a graphical model. Then, once it is established that Δ is conformal, the next step is to check that it is decomposable.

The first task can be carried out using the following result.

Lemma 5.10. *A hypergraph $\mathcal{H}$ is not graphical if and only if $\exists h \in \mathcal{H}$ and $k \in \{1, \ldots, K\}$ such that $\{k\} \cup h$ is complete in $\mathcal{G}(\mathcal{H})$.*

Proof. For any graph G, let $\mathcal{C}(G)$ be its clique hypergraph, the hypergraph whose hyperedges are the cliques of G. By definition, $\mathcal{H}$ is graphical whenever $\mathcal{H} = \mathcal{C}(\mathcal{G}(\mathcal{H}))$, where $\mathcal{G}(\mathcal{H})$ is the interaction graph of $\mathcal{H}$. By construction, $\mathcal{H} \sqsubseteq \mathcal{C}(\mathcal{G}(\mathcal{H}))$ so that the only case in which a hypergraph $\mathcal{H}$ is not graphical is when there exists a subset $s \subset \mathcal{K}$ which is complete in $\mathcal{G}(\mathcal{H})$ and $s \supsetneq h$ for some $h \in \mathcal{H}$. ■

The pseudo-code for the corresponding algorithm is given in Table 12. There are other alternatives to Lemma 5.10; see, for example, Berge (1989) and Badsberg (1995).

Finally, in order to decide whether a graphical model Δ is decomposable, a combination of the Maximum Cardinality Search Algorithm (see, for example, Golumbic, 2004) and Algorithm 4.11 from Cowell et al. (2003), can be used to achieved the following:

1. Verify that Δ is decomposable by checking whether a perfect numbering exists.
2. For decomposable Δ, produce a perfect sequence of its cliques.

The pseudo-code is given in Table 13. An almost identical algorithm, capable of completing both tasks as well, is also provided by Blair and Barry (1993, Section 4.2.1).

5.3 Table Collapsing and the Extended MLE

The factorization results of this section, summarized by Theorem 5.3, permits to generalize to the extended log-linear modeling framework the collapsibility conditions for contingency tables given by Asmussen and Edwards (1983).

Let Δ be a hierarchical log-linear model, where Δ, according to the conventions established here, is identified with the hypergraph of its facets. A hypergraph is called *simple* if, trivially, it has one hyperedge. Following Asmussen and Edwards (1983), given a log-linear model Δ and a non-empty subset of factors $a \in 2^{\mathcal{K}}$, let $\Delta_a = \Delta \wedge \{a\}$, where $\{a\}$ denotes the simple hypergraph induced by a and $\mathcal{I}_a = \bigotimes_{k \in a} \mathcal{I}_k$.

Example 5.11. For $\Delta = [12][234][14]$ and $a = \{1, 2, 3\}$, $\Delta_a = [12][23]$. Example borrowed from Asmussen and Edwards (1983, Section 2). ■

Definition 5.12. The model Δ is collapsible into $a \in 2^{\mathcal{K}}$ if $\widehat{\mathbf{m}}(i_a) = \widehat{\mathbf{m}}^a(i_a)$, for all $i_a \in \mathcal{I}_a$.

In the above definition, the extended MLE is assumed, an improvement over the characterization of collapsibility offered by Asmussen and Edwards (1983), in which no attention to the existence of the MLE[2] is paid.

Theorem 5.13. *A model Δ can be collapsed into $a \in 2^{\mathcal{K}}$ if and only if Δ is reducible and Δ_a is a decomposing sub-complex of Δ.*

Proof (sketched). If the conditions of the theorem are satisfied, it follows from Equation (79) of Theorem 5.3 that, by summing over the cells in a^c, the terms in the numerator associated for the decomposing sub-complexes different than Δ_a cancel out with the margins in the numerator representing the separators of the decomposition, for each $i \in \mathcal{I}$.

[2] It should be remarked that Asmussen and Edwards (1983)'s proof of the necessity of the conditions of collapsibility implicitly assumes an extended log-linear modeling framework, for it entails the construction of a model in which some coefficients in the log-linear expansion are let to be minus infinity.

As for necessity, assume that the table is collapsible. Then, each connected component of $\partial(a^c)$ in $\mathcal{G}(\Delta)$ is contained in a generating class $d \in \Delta$, which is precisely the condition of Theorem 2.3 of Asmussen and Edwards (1983). To this end, let C_a the matrix representing the linear operator for summing over the cell in $a^c = \mathcal{K} \setminus a$ and let S_a be the analogous matrix for summing over the marginal configurations in a^c. Suppose, arguing by contradiction, that, for some connected component b of a^c, $\partial b \not\subseteq d$ for all $d \in \Delta$. Then ∂b corresponds to a marginal configuration different than any of the marginal configurations associated to Δ_a. As a result,

$$\mathrm{S}_a\mathrm{A}_\Delta\widehat{\mathbf{m}} = \mathrm{S}_a\mathrm{A}_\Delta\mathbf{n} \neq \mathrm{A}_{\Delta_a}\widehat{\mathbf{m}}^a,$$

where the characterization of the extended MLE as the only vector in the toric variety which satisfies the moment equations is used (see Theorem 2.6). However, by hypothesis, $\mathrm{A}_{\Delta_a}\widehat{\mathbf{m}}^a = \mathrm{A}_{\Delta_a}\mathrm{C}_a\widehat{\mathbf{m}}$. Hence $\mathrm{S}_a\mathrm{A}_\Delta\widehat{\mathbf{m}} \neq \mathrm{A}_{\Delta_a}\mathrm{C}_a\widehat{\mathbf{m}}$, which is a contradiction since $\mathrm{S}_a\mathrm{A}_\Delta$ and $\mathrm{A}_{\Delta_a}\mathrm{C}_a$ gives, by assumption and in virtue of the moment equations, the same marginal sums. The existence of a decomposing sub-complex Δ_a follows then from the decomposition arguments in the first part of Theorem 2.3 of Asmussen and Edwards (1983). ■

Remarks.

1. As an alternative proof of Theorem 5.13, it is possible to show that the conditions of Theorem 5.13 are equivalent to the conditions of Theorem 2.3 of Asmussen and Edwards (1983), i.e. that each connected component of $\partial(a^c)$ in $\mathcal{G}(\Delta)$ is contained in a facet of Δ. This equivalence can be proved using the fact that two subsets a and b of $\mathcal{K}$ induce a decomposition of Δ if and only if $a \cap b \subseteq d$ for some $d \in \Delta$ and $a \cap b$ separates $a \setminus b$ and $b \setminus a$ in $\mathcal{G}(\Delta)$ (see, for example, Lauritzen, 2002). Furthermore, the induced decomposition has $\Delta_a = \Delta \wedge \{a\}$ and $\Delta_b = \Delta \wedge \{b\}$ as decomposing sub-complexes.

2. The graphical criteria for collapsibility given by Asmussen and Edwards (1983) are easier to verify than the ones offered here. Theorem 5.13 unravels the connections between collapsibility and reducibility of a hierarchical log-linear model and shows that they hold unchanged also for the extended MLE.

Example 5.14 (Collapsible models).

1. Collapsing is not possible in Example 5.11, since $\Delta_a = [12][23]$ is not a decomposing sub-complex of $\Delta = [12][234][14]$.

2. Let $\Delta = [13][124][235]$ and $a = \{4, 5\}$, as in Example 2 of Asmussen and Edwards (1983). Then $\Delta_a = [12][13][23]$. Decompose Δ into $\Delta_1 = [235]$ and $\Delta_2 = [13][124][23]$. Next decompose Δ_2 into $\Delta_3 = [124]$ and $\Delta_4 = [12][13][23]$. The decomposing sub-complexes are

$$\Delta_1 = [235], \quad \Delta_3 = [124] \quad \text{and} \quad \Delta_4 = [12][13][23].$$

   Since $\Delta_4 = \Delta_a$, the model is collapsible into a.

3. Let $\Delta = [123][234][1245]$ and $a = \{1, 2, 3, 4\}$, the opening example of Bishop et al. (1975, Section 3.7.3: *Fitting by Stages*). Then $\Delta_a = [123][234][124]$, while Δ can be decomposed into $\Delta_1 = [1245]$ and $\Delta_2 = [123][234][124]$. Since $\Delta_2 = \Delta_a$, the model is collapsible into a. The other example from Section 3.7.3 of Bishop et al. (1975) was already shown in Example 5.4 to be reducible with decomposing sub-complexes $[123][234][134]$ and $[5]$, the former being Δ_a for $a = \{1, 2, 3, 4\}$, so the model is collapsible into a.

4. The simplest example of non-collapsibility is the model $\Delta = [12][23]$ with $a = \{1, 3\}$, from which it follows $\Delta_a = [1][3]$, which is not a decomposing sub-complex of Δ. ■

6 Testing for Goodness of Fit

Extended MLEs can be used to perform goodness-of-fit tests and, more importantly, model selection in a very similar way to the ordinary MLEs. In fact, it is a well known fact that, when the MLE exists, the asymptotic distribution of any member of the power divergence family of Cressie and Read (1988), including as special cases Pearson's χ^2 and the likelihood ratio statistics, is χ^2 with a number of degrees equal to the difference between the number of cells $|\mathcal{I}|$ and the dimension of the log-linear subspace $\mathcal{M}$ or, equivalently, of the marginal cone. When the MLE is not defined, the χ^2 approximation fails.

Within the extended exponential family framework described in Rinaldo (2006), nonexistence of the MLE is associated to a reduced exponential family of distribution for the cell counts (the "boundary" log-linear model) supported on the facial set $\mathcal{F}$ corresponding to the face F of marginal cone whose relative interior contains the observed sufficient statistics. With respect to this reduced family, the asymptotics for the goodness-of-fit statistics is identical to the case in which the MLE is defined, except that the likelihood zeros are treated as structural zeros, hence not affecting the likelihood.

In practice, this entails replacing the MLE with the extended MLE and adjusting the number of degrees of freedom, which are now to be computed as the difference between for the cardinality of the facial set $|\mathcal{F}|$, i.e. the number of cell mean values that can be estimated, and the number of parameters in the restricted model, namely $\dim(\mathcal{M}_{\mathcal{F}})$ or, equivalently, $\dim(F)$.

Therefore, when boundary log-linear models are allowed, testing for a specific log-linear model can be done when the MLE is nonexistent and, provided that a procedure to calculate the appropriate facial set and the extended MLE is available, is a relatively straightforward task. Below, we illustrate by means of examples various practical aspects of goodness-of-fit testing when the MLE is nonexistent and it is required to adjust the number of degrees of freedom in order to obtain meaningful tests.

Example 6.1. The following pattern of zeros for the 2^3 table and the model $\Delta = [12][13][23]$ of no-second-order interaction, due to Haberman (1974), has been for a long time the only known instance of non-existent MLE with positive margins:

0	

	0

.

The two zeros are in fact likelihood zero, exposing one of the 16 facets of the corresponding marginal cone. The dimension of the log-linear subspace for this model, or, equivalently, of the marginal cone, is 7, leaving 1 degree of freedom when the MLE exists. However, because of the likelihood zeros, inference can only be made for the 6-dimensional exposed facet. Since the cardinality of the associated facial set $\mathcal{F}$ is also 6, the resulting boundary log-linear model is the saturated model on $\mathcal{F}$, so the correct number of degrees of freedom is 0, and the χ^2asymptotic approximation for goodness-of-fit statistics cannot be applied.

Example 6.2. The patters of zeros in the 3^3 table below form a set of likelihood zeros for the model $\Delta = [12][13][23]$ because they expose one of the 207 facets of the corresponding marginal cone:

0		
0	0	

	0	0
	0	

0		0
		0

.

The dimension of the reduced boundary model, i.e. $\dim(\mathcal{M}_{\mathcal{F}})$, is 18, which is also the cardinality of the facial set for this configuration of likelihood zeros. As in the previous example, this defines the saturated model on $\mathcal{F}$, giving 0 degrees of freedom and making χ^2 approximations not applicable.

Example 6.3. Under the same log-linear model $\Delta = [12][23][13]$, the pattern of likelihood zeros

0		
0		

0		
0		

	0	0

will imply that the number of degrees of freedom for the χ^2 test is 3, because the total number of estimable cell mean values is 21 and the number of parameters for the reduced model corresponding to the facet defined by the zeros is 18. ■

Example 6.4. If instead the likelihood zeros correspond to a null margin, like in the table below, then, by the same token, the adjusted number of degrees of freedom is 6, obtained by subtracting 18, the dimension of the facet, from 24, the cardinality of the facial set.

0		
0		
0		

■

As pointed out by Fienberg and Rinaldo (2006), the R (R Development Core Team, 2005) routines `loglin` and `glm`, as well as virtually any other software for inference and model selection for log-linear models, do no detect such degeneracy and report the un-adjusted, incorrect numbers of degrees of freedom for all the examples below. It is important to remark on the combined, misleading consequences of the incorrect calculation of the degrees of freedom with the habit of adding small positive quantities to the zero cells. This mi-practice, which is commonly observed in applications involving sparse tables and implemented in many statistical software, such as SAS, is thought to facilitate the convergence of the underlying numerical procedure for computing the MLE. For the tables in Examples 6.1, 6.2 and 6.3, this modification will produce an erroneous MLE that will be extremely close to the original table and, ultimately, will make any measure of goodness-of-fit almost zero. Since the un-adjusted number of degrees of freedom for this model, assuming an existing MLE, is 1 for the table in Example 6.1 and 8 for the ones in Examples 6.2 and 6.3, it is virtually almost certain that the null hypothesis will not be rejected, *no matter what the correct model is*.

We conclude this section with a real life example from genetics, borrowed from (Edwards, 2000, Section 2.2.5). See again Fienberg and Rinaldo (2006) for another real life example involving clinical trials.

Example 6.5. The dataset reproduced in Table 4 is a sparse 2^6 contingency table which was obtained from the cross-classification of six dichotomous categorical variables, labeled with the letters

			1				2				**E**
			1		2		1		2		**C**
			1	2	1	2	1	2	1	2	**A**
1	1	1	**0**	16	**0**	1	**0**	4	**0**	1	
		2	1	**0**	**0**	**0**	1	**0**	**0**	**0**	
	2	1	3	2	**0**	**0**	1	**0**	**0**	**0**	
		2	7	**0**	1	**0**	4	**0**	**0**	**0**	
2	1	1	**0**	1	1	4	**0**	**0**	**0**	4	
		2	**0**	**0**	**0**	**0**	**0**	**0**	2	**0**	
	2	1	**0**	**0**	1	**0**	**0**	**0**	**0**	1	
		2	1	**0**	3	**0**	**0**	**0**	11	**0**	
F	**D**	**B**									

Table 4: Cell counts for the chromosome mapping example from (Edwards, 2000, Section 2.2.5). Data publicly available on the MIM website: http://www.hypergraph.dk/.

A-F, recording the parental alleles corresponding to six loci along a chromosome strand of a barely powder mildew fungus, for a total of 70 offspring. The data were originally described by Christiansen and Giese (1991) and further analyzed by Edwards (1992).

Edwards (2000) utilizes this table to illustrate how MIM, the companion software to his book, is capable of performing automated model selection procedure for graphical models. MIM implements the Modified Iterative Proportional Scaling (MIPS) algorithm proposed by Frydenberg and Edwards (1989) for fitting general hierarchical interaction models (see also Lauritzen, 1996, Section 6.4.3). The MIPS algorithm performs a sequence of cyclic optimizations of the likelihood function along line sections of the natural parameter space corresponding to the coordinates of the canonical sufficient statistic, as described, for example, by Lauritzen (1996, Section D.1.5). When the sufficient statistic is a cut (see Bardorff-Nielsen, 1978, page 50), the MIPS takes advantage of the consequent factorization of the joint density by performing separate maximizations of the distinct, variation independent, factors of the density. As a result, for decomposable models, MIPS is essentially identical to the IPF algorithm.

In the example of Table 4, MIM searches for the optimal graphical model using a backward selection approach, starting from the saturated model and then testing for the removal of each edge, one at a time. The test statistic is Pearson's χ^2. The sequence of nested decomposable models selected by MIM is shown in Table 5 along with the degrees of freedom used by MIM, in the second column. The final model, depicted in Figure 5, is consistent with the expectation, motivated by the biological theory, of no interference.

Due to the high sparsity level of the table, nonexistence of the MLE affects at all levels of the

Model	reported d.f.	correct d.f.
[ABCDEF]	0	0
[ABCEF] [ABCDE]	16	3
[BCEF] [ABCDE]	24	6
[BCEF] [ABCE] [ABCD]	32	12
[BCEF] [ABCE] [ABD]	36	17
[BCEF] [AD] [ABCE]	38	18
[CEF] [AD] [ABCE]	42	22
[CEF] [AD] [BCE] [ABE]	46	27
[CEF] [AD] [ABE]	48	29
[CEF] [AD] [BE] [AB]	50	31
[CF][CE][AD] [BE] [AB]	52	37

Table 5: Hierarchy of nested decomposable models fitted by `MIM` on the dataset of Table 4, starting from the saturated model on top down to the selected model, depicted in Figure 5. The marginal configurations containing zero entries are shaded. The second column shows the incorrect numbers of degrees of freedom reported by `MIM` and the third column the adjusted degrees of freedom after accounting for the non-estimable parameters.

hierarchy of nested decomposable sub-models fitted by `MIM`. This is particularly easy to spot in this example because all fitted models are decomposable, the only case in which nonexistence is completely characterized by null margins (see Section 5.2). The marginal configurations containing null terms are shaded in color in Table 5. Despite claiming to perform adjustment for sparsity in the computation of the degrees of freedom, `MIM` fails to detect all these cases of nonexistence, as indicated by the fact that all the tested models are reported to have the numbers of degrees of freedom that would be appropriate should the MLE exist. The correct number of degrees of freedom, obtained by counting the number of estimable cells minus the number of parameters for the corresponding restricted models, is reported in the third column. Note that the discrepancies between the two columns are rather significant.

For this example, the reason why nonexistence goes undetected is essentially due to the decomposable nature of these models and the fact that the MIPS algorithm behaves exactly like the

IPF algorithm. As proved in Section 5.2, for decomposable models the MLEs can be factorized into rational functions of the margins, in such a way that a null margin will imply a vector of fitted values with zero entries corresponding exactly to that margins. From the algorithmic point of view, IPF, by design, replicates this property[3] and, moreover, converges almost immediately. As a result, for decomposable models, nonexistence of the MLE will not produce any of the numerical inconveniences that can be observed for other types of models, like in the examples of Fienberg and Rinaldo (2006).

As a final remark, in the original analysis performed by Edwards (2000) the sparsity of the sample is taken into account by using exact tests based on complete enumerations (Edwards, 2000, Section 5.4) rather than the χ^2 asymptotic approximations to the test statistic. Since degrees of freedom considerations do not apply to this type of tests and the models considered are all decomposable (see Section 5.2), the procedure followed by Edwards (2000) is to be considered essentially correct, as least as far as existence of the MLE is concerned. (Of course, it remains to be proved whether exact tests for log-linear models are optimal and in which sense but this is a separate, still unresolved issue.) The sequence of nested decomposable models obtained using exact test is slightly different than the one displayed in Table 4, but the final model is the same one. ■

Figure 5: Optimal model for the Table 4 determined by `MIM` using the backward selection procedure. Graphical output produced by `MIM`.

[3]Recall that, for decomposable models, the IPF procedures will compute both the MLE and the extended MLE in exactly the same way.

7 Tables of Pseudo-Codes

```
0   : function MultiIndex_to_Index( [i_1,...,i_K],[I_1,...,I_K] )
1   : index = 1 + (i_K − 1)
2   : for j = 1 to K − 1
2.1 :     index = index + i_j * ( ∏_{l=j+1}^{n} I_l )
3   : end
4   : return index
```

Table 6: `MultiIndex_to_Index`. Pseudo-code for linearizing cell label combinations in a lexicographic order, according to the function indicated in Equation (1). It requires two inputs, the multi-index vector $i = [i_1, \ldots, i_n]$ and a second vector $[I_1, \ldots, I_K]$, whose elements are ordered accordingly to i, holding the number of levels for the corresponding categories. The output is an integer index. See also the inverse function `Index_to_MultiIndex` in Table 7.

```
0     : function Index_to_MultiIndex( index , [I_1,...,I_n] )
1     : for i = 1 to n
1.1   :     if i == n
1.1.1 :         multi_index[i] = index
1.2   :     else
1.2.1 :         multi_index[i] = floor( (index - 1)/(∏_{j=i+1}^{K} I_j) ) + 1
1.2.2 :         index = index - (∏_{j=i+1}^{K} I_j) *(multi_index[i] - 1)
1.2   :     end
2     : end
3     : return multi_index
```

Table 7: `Index_to_MultiIndex`. Pseudo-code for the inverse function of the bijection in Equation (1) (see the pseudo-code for `MultiIndex_to_Index` in Table 6). It takes as input an integer `index` and an order list of holding the number of levels for the categories $[I_1, \ldots, I_n]$. The output is a vector `multi_index` such that `index = MultiIndex_to_Index( multi_index , [`$I_1, \ldots, I_n$`] )`.

```
1          : nz = 0
2          : for binstr = 0 to 2^|h| − 1
2.1        :     nz = nz + 1
2.2        :     coord[nz] = 1
2.3        :     v[nz] = 1
2.4        :     for b = |h| to 1 by -1
2.4.1      :         bv = value of b-th bit of binstr
2.4.2      :         if bv = 0
2.4.2.1    :             if i_{k_b} = 1
2.4.2.1.1  :                 v[nz] = 0
2.4.2.1.2  :                 leave 2.4
2.4.2.2    :             end
2.4.2.3    :             v[nz] = -v[nz]
2.4.2.4    :             coord[nz] = coord[nz] + Ī_{k_b}(i_{k_b} − 2)
2.4.3      :         else
2.4.3.1    :             if i_{k_b} = I_{k_b}
2.4.3.1.1  :                 v[nz] = 0
2.4.3.1.2  :                 leave 2.4
2.4.3.2    :             end
2.4.3.3    :             coord[nz] = coord[nz] + ∏_{l>b}^{|h|} I_{k_l}(i_{k_b} − 1)
2.4.4      :         end
2.5        :     end
2.6        :     if v[nz] = 0
2.6.1      :         nz = nz -1
2.7        :     end
3          : end
4          : return nz, v, coord
```

Table 8: Pseudo-code of the algorithm to compute the values of nonzero entries of the U_h^k matrix associated to a cell $< i_1, \ldots, i_K >$, with the matrix U_k^h defined using Z_k as in (26). It returns nz, the number of nonzero entries in the row and the two vectors coord and v, both of length nz, containing the coordinates and values of the nonzero elements, respectively. The symbol $\bar{I}_{k_b} = \prod_{l>b}^{|h|} I_{k_l}$ if $b < I_{k_b}$ and $\bar{I}_{k_b} = 1$ if $b = I_{k_b}$. The value of bv in line 2.4.1 can be easily obtained by setting bs = binstr outside the loop starting at 2.4 and then replacing 2.4.1 with the two statements: bv = mod(bs,2) and bs = floor(bs/2). See Section 3.4.2.

```
1     : v = 1
2     : for b = |h| to 1 by -1
2.1   :     if i_{k_b} = 1
2.1.1 :         l[b] = {1,2,...,I_{k_b} − 1}
2.2   :     else
2.2.1 :         v = -v
2.2.2 :         l[b] = i_{k_b} − 1
2.3   :     end
3     : end
4     : nz = 0
5     : L = l[1] x l[2] x ...  x l[|h|]
6     : for each multi_index in L
6.1   :     nz = nz + 1
6.2   :     coord[nz] = MultiIndex_to_Index( multi_index , [I_{k_1},...,I_{k_|h|}] )
7     : end
8     : return nz, v, coord
```

Table 9: Pseudo-code of the algorithm to compute the values of nonzero entries of the U_h^k matrix associated to a cell $< i_1, \ldots, i_K >$, with the matrix U_k^h defined using C_k as in (28). It returns `nz`, the number of nonzero entries in the row, their coordinates `coord` and their unique values `v`. The length of the vector is `nz`. For each `b` the entry `l[b]` of the list `l` is either an ordered list of numbers, as in line `2.1.1`, or a scalar. The function `MultiIndextoIndex` from Table 6 is used in line `6.2`. See Section 3.4.2.

```
1 : ih = π_h(i_1,...,i_K)
2 : Ih = π_h(I_1,...,I_K)
3 : coord = MultiIndex_to_Index( ih , Ih )
4 : return coord
```

Table 10: Pseudo-code of the algorithm to compute the nonzero entry `coord` of the W_h matrix associated to a cell $< i_1, \ldots, i_K >$ from Section 3.4.2. The coordinate projection function $\pi_h : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}_h}$ is defined consistently with the notation of Section 1.1: $\pi_h(\mathbf{x}) = \{x_k \colon k \in h\}$ for any $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$. The function `MultiIndextoIndex` from Table 6 is used in line 3.

```
1     : coord = 0
2     : for b = 1 to |h|
2.1   :     if i_{k_b} = I_{k_b}
2.1.1 :         coord = -1
2.1.2 :         break 2
2.2   :     end
2.3   :     coord = coord * (I_{k_b} − 1) + (i_{k_b} − 1)
3     : end
4     : coord = coord + 1
5     : return coord
```

Table 11: Pseudo-code of the algorithm to compute the nonzero entry `coord` of the V_h matrix associated to a cell $< i_1, \ldots, i_K >$. See Section 3.4.2.

```
1       : graphical = yes
2       : G[ 1...,n , 1,...,n ] = 0
3       : for i = 1 to p
3.1     :     G[ h_i , h_i ] = 1
4       : end
5       : for i = 1 to p
5.1     :     h_i^c = {1,...,K}\h_i
5.2     :     for j = 1 to |h_i^c|
5.2.1   :         if G[ h_i ∪ h_i^c(j) , h_i ∪ h_i^c(j) ] = E_{|h_i|+1}
5.2.1.1 :             graphical = no
5.2.1.2 :             return graphical
5.2.2   :         end
5.3     :     end
6       : end
7       : return graphical
```

Table 12: Pseudo-code for checking whether a hypergraph $\mathcal{H} = \{h_1, \ldots, h_p\}$ on n nodes is conformal, using Lemma 5.10. The algorithm builds and utilizes the $n \times n$ incidence matrix `G` whose (i,j)-th entry is 1 if $\{i,j\} \subseteq h$ for some $h \in \mathcal{H}$ and 0 otherwise. For a positive integer k, E_k denotes here the $k \times k$ matrix whose entries are all 1's. The notation $h_i(j)$ means the j-th entry of the hyperedge h_i. See Section 5.2.3.

```
1         : decomposable = yes
2         : i = l = 1
3         : L = ∅
4         : w[n] = 0, n = 1,...,K
5         : numbering[n] = 0, n = 1,...,K
6         : sequence = ∅
7         : while( L ≠ V )
7.1       :     U = V\L
7.2       :     j = { argmax w[n] :  n ∈ U }
7.3       :     numbering[j] = i
7.3       :     Π_i = ne(j) ∩ L
7.4       :     if Π_i is not complete in G
7.4.1     :         decomposable = no
7.4.2     :         return decomposable
7.5       :     else
7.5.1     :         w[n] = w[n] + 1, n ∈ ne(j) ∩ U
7.5.2     :         if i >1
7.5.2.1   :             if |Π_i| < 1 + |Π_{i-1}| or i = K
7.5.2.1.2 :                 C_l = j ∪ Π_i
7.5.2.1.2 :                 l = l + 1
7.5.2.2   :             end
7.5.3     :         end
7.6       :     end
7.7       :     L = L ∪ {j}
7.8       :     i = i + 1
8         : end
9         : return (C_1,...,C_p)
```

Table 13: Pseudo-code for the variation of the Maximum Cardinality Search Algorithm mentioned in Section 5.2.3. The input is the interaction graph G derived from the graphical hierarchical log-linear model Δ with p facets. If Δ is not decomposable, the routine will return the flag `decomposable = no`. Otherwise, it will produce an ordered sequence $(C_1, \ldots, C_p)$ of cliques of Δ which form a perfect sequence for the associated interaction graph.

```
1   : κ_x(j) = 0  j = 1,...,r
2   : for i = 1 to |I|
2.1 :     determine j such that χ_j(i) = 1
2.2 :     κ_x(j) = κ_x(j) + β_x(i)
3   : end
4   : return κ_x
```

Table 14: Pseudo-code for computing $\boldsymbol{\kappa}_{\mathbf{x}}$. See the end of Section 4.3.2.

```
1   : Grad[i] = 0  i = 1,...,|I|
2   : for i = 1 to |I|
2.1 :     determine j such that χ_j(i) = 1
2.2 :     Grad[i] = n(i) - (N_j / κ_x(j)) b_x(i)
3   : end
4   : return Grad
```

Table 15: Pseudo-code for computing $\nabla \ell_{\mathcal{L}}$. It is assumed that the vector $\boldsymbol{\kappa}_{\mathbf{x}}$ has been pre-computed. See the end of Section 4.3.2.

```
1     : Hess = 0
2     : for j = 1 to r
2.1   :     w = 0
2.2   :     loop for i ∈ supp(χ_j)
2.2.1 :         compute v_i^T  (i-th row of V)
2.2.2 :         Hess = Hess - (N_j / κ_x(j)) v_i v_i^T
2.2.3 :         w = w + b_x(i) v_i^T
2.3   :     end
2.4   :     Hess = Hess + (N_j / κ_x^2(j)) w w^T
3     : end
4     : return Hess
```

Table 16: Pseudo-code for computing the hessian $\nabla^2 \ell_{\mathcal{L}}(\mathbf{x})$. It is assumed that the vector $\boldsymbol{\kappa}_{\mathbf{x}}$ has been pre-computed. See the end of Section 4.3.2.

```
1       : out[i] = 0, i = 1,...,I
2       : for i = 1 to I
2.1     :     [ coord, nz, v ] = Get_Row( i )
2.2     :     if nz > 0
2.2.1   :         for k = 1 to nz
2.2.1.1 :             j = coord[k]
2.2.1.2 :             out[i] = out[i] + x_j * v[k]
2.2.2   :         end
2.3     :     end
3       : end
4       : return out
```

Table 17: Pseudo-code for computing $\mathrm{U}\mathbf{x}$. See Section 4.3.4. Note that, using the sparse basis representation V_h from Section 3.4.2, `v[k] = 1`, so the statement `2.2.1.2` becomes `out[i] = out[i] + x_j`.

```
1       : out[j] = 0, j = 1,...,k
2       : for i = 1 to I
2.1     :     [ coord, nz, v ] = Get_Row( i )
2.2     :     if nz > 0
2.2.1   :         for k = 1 to nz
2.2.1.1 :             j = coord[k]
2.2.1.2 :             out[j] = out[j] + y_i * v[k]
2.2.2   :         end
2.3     :     end
3       : end
4       : return out
```

Table 18: Pseudo-code for computing $\mathrm{U}^\top\mathbf{y}$. See Section 4.3.4. Note that, using the sparse basis representation V_h from Section 3.4.2, `v[k] = 1`, so statement `2.2.1.2` becomes `out[j] = out[j] + y_i`.

```
1           : A = 0
2           : for i = 1 to I
2.1         :     [ coord, nz, v ] = Get_Row( i )
2.2         :     if nz > 0
2.2.1       :         for j_index = 1 to nz
2.2.1.1     :             j = coord[j_index]
2.2.1.1.1   :                 for k_index = 1 to j_index
2.2.1.1.1.1 :                     k = coord[k_index]
2.2.1.1.1.2 :                     A[k,j] = A[k,j] + x_i * v[j_index] * v[k_index]
2.2.1.1.2   :                 end
2.2.1.2     :             end
2.2.2       :         end
2.3         :     end
3           : end
```

Table 19: Pseudo-code for computing $\mathrm{U}^\top \mathrm{D}_\mathbf{x} \mathrm{U}$. See Section 4.3.4.

```
1   : [ coord, nz ] = Get_Row( [i_1,...,i_K] )
2   : l = coord[1]
3   : v_i = -b_l
4   : for k = 2 to nz
4.1 :     j = coord[k] - r
4.2 :     v_i(j) = v_i(j) + u_i^(2)(k)
5   : end
5   : return v_i
```

Table 20: Pseudo-code for generating the row of the matrix V associated to the cell $< i_1, \ldots, i_K >$ according to the results of Section 4.3.5. It uses the function `Get_Row` (as described in Section 4.3.4) and the matrix U_2.

Acknowledgments

This documents describes results extracted from my Ph.D. dissertation work, completed under the supervision of Stephen E. Fienberg, to whom I am thankful for guidance and advice. Sections 3.2, 3.4 and 7, present some unpublished material from Fienberg et al. (1980).

References

Agresti, A. (2002). *Categorical Data Analysis* (second ed.). New York: John Wiley & Sons.

Asmussen, S. and D. Edwards (1983). Collapsibility and response variables in contingency tables. *Biometrika 70*, 567–578.

Badsberg, J. H. (1995). *An environment for graphical models,*. Ph. D. thesis, Department of Mathematics and Computer Science, Aalborg University.

Bailey, R. A. (2004). *Association Schemes : Designed Experiments, Algebra and Combinatorics*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Bardorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. New York: John Wiley & Sons.

Berge, C. (1989). *Hypergraphs*. North-Holland Mathematical Library. Elsevier Science Pub Co.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society 25*, 220–233.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge, Massachussetts: MIT Press.

Björner, A., M. Las Vergnas, B. Sturmfels, N. White, and G. M. Ziegler (1999). *Oriented Matroids* (Second ed.). Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Blair, J. S. R. and P. Barry (1993). An introduction to chordal ggraphs and clique trees. In A. George, G. J. R., and J. W. H. Liiu (Eds.), *Graph Theory and Sparse Matrix Computation*, Number 56 in The IMA Volumes in Mathematics and Its Applications. Springer Verlag.

Borwein, J. M. and A. S. Lewis (2000). *Convex Analysis and Nonlinear Optimization : Theory and Examples* (Second ed.). CMS Books in Mathematics. Springer.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization* (second ed.). Cambridge University Press.

Christiansen, S. K. and H. Giese (1991). Genetic analysis of obligate barley powdery mildew fungus based on rfpl and virulence loci. *Theoretical and Applied Genetics 79*, 705–712.

Cowell, R., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (2003). *Probabilistic Networks and Expert Systems*. Information Science and Statistics. Springer Verlag.

Cox, D., J. Little, and D. O'Sheas (1996). *Ideals, Varieties and Algorithms. An Introduction to Computational Algebraic Geometry and Commutative Algebra* (second ed.). Springer-Verlag.

Cressie, N. A. C. and T. R. C. Read (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability 3*(1), 146–158.

Darroch, J. and T. Speed (1983). Additive and multiplicative models and interactions. *The Annals of Statistics 11*, 724–738.

Darroch, J. N. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics 43*, 1470–1480.

De Loera, J. and S. Onn (2006). Markov bases of 3-way tables are arbitrarily complicated. *Journal of Symbolic Computation 41*, 173–181.

Deming, W. E. and F. F. Stephan (1940). On a least square adjustment of a sampled ffrequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*, 427–444.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Volume 11 of *Lecture Notes*. Institue of Mathematical Statistics.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics2 26*(1), 363–397.

Dobra, A. . and S. E. Fienberg (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences 97*(22), 11885–11892.

Edwards, D. (1992). Linkage analysis using log-linear models. *Computational Statistics and Data Analysis 13*, 281–290.

Edwards, D. (2000). *Introduction to Graphical Modelling* (second ed.). Springer.

Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant (2006). Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computations 41*, 222–233.

Fienberg, S. E. (1980). *The analysis of Cross-Classified Categorical Data* (Second ed.). MIT Press.

Fienberg, S. E. and J. P. Gilbert (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association 65*(330), 694–701.

Fienberg, S. E., M. Meyer, and G. W. Stewart (1980). The numerical analysis of contingency tables. Unpublished manuscript.

Fienberg, S. E. and A. Rinaldo (2006). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. Accepted in Journal of Statistical Planning and Inference.

Forster, J. J. (2003). Bayesian inference for poisson and multinomial log-linear models. Technical report, School of Mathematics, University of Southampton. http://www.maths.soton.ac.uk/staff/JJForster/Papers/MultPoiss.pdf.

Frydenberg, M. and D. Edwards (1989). A modified iterative scaling algorithm for estimation in regular exponential families. *Computational Statistics and Data Analysis 8*, 142–153.

Fulton, W. (1978). *Introduction to toric varieties*. Pinceton: Princeton University Press.

Gawrilow, E.and Joswig, M. (2000). Polymake: a framework for analyzing convex polytopes. In *Polytopes, Combinatorics and Computation*. Boston, Massachusetts: Birkhauser.

Geiger, D., C. Meek, and B. Sturmfels (2006). On the toric algebra of graphical models. To appaer in the *Annals of Statistics*. Available at http://www.research.microsoft.com.

Geng, Z. (1989). Decomposability and collapsibility for log-linear models. *Journal of the Royal Statistical Association 38*, 189–197. Ser. C.

Gloneck, G., J. N. Darroch, and T. P. Speed (1988). On the existence of the maximum likelihood estimator for hierarchical log-linear models. *Scandinavia Journal of Statistics 15*, 187–193.

Golumbic, M. C. (2004). *Algorithmic Graph Theory and Perfect Graphs* (Second ed.). Number 57 in The Annals of Discrete Mathematics. Elsevier Science.

Grünbaum (2003). *Convex Polytopes* (Second ed.). New York: Springer-Verlag.

Haberman (1974). *The Analysis of Frequency Data*. Chicago, Illinois: University of Chicago Press.

Haberman, S. J. (1972). Log-linear fit for contingency tables - Algorithm AS51. *Applied Statistics 21*, 218–225.

Haberman, S. J. (1976). Correction to as 51: Log-linear fit for contingency tables. *Applied Statistics 25*(2), 193.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected counts. *The Annals of Statistics 5*(6), 1148–1169.

Hemmecke, R. and R. Hemmecke (2003, September). 4ti2 version 1.1—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. www.4ti2.de.

Hosten, S. and S. Sullivant (2004). Ideals of adjacent minor. *Journal of Algebra 277*, 615–642.

Knuiman, M. W. and T. P. Speed (1988). Incorporating prior information into the analysis of contingency tables. *Biomatrics 44*, 1061–1071.

Lang, J. B. (2004). Multinomial-Poisson Homogeneous models for contingency tables. *The Annals of Statistics 32*(1), 430–383.

Lauritzen, S. F. (1996). *Graphical Models*. New York: Oxford University Press.

Lauritzen, S. L. (2002). Lectures on contingency tables. electronic edition. http://www.math.aau.dk/~steffen/cont.pdf.

Leimer, H.-G. (1993). Optimal decomposition by clique separators. *The Annals of Discrete Mathematics 113*, 99–123.

Oda, T. (1988). *Convex Bodies and Algebraic Geometry: An Introduction to the Theory of Toric Varieties*. Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 3 Folge. Springer.

Pachter, L. and B. Sturmfels (Eds.) (2005). *Algebraic Statistics for Computational Biology*. Cambridge University Press.

Pistone, G., E. Riccomagno, and W. P. Wynn (2000). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall/CRC.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rinaldo, A. (2005). *Maximum Likelihood Estimates in Large Sparse Contingency Tables*. Ph. D. thesis, Carnegie Mellon University, Department of Statistics.

Rinaldo, A. (2006). On maximum likelihood estimation for log-linear models. Technical Report 833, Department of Statistics, Carnegie Mellon University.

Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *The Annals of Statistics 23*(4), 1160–1174.

Schrijver, A. (1998). *Theory of Integer and Linear Programming*. New York: John Wiley & Sons.

Serre, J. P. (1977). *Linear Representations of Finite Groups*. Springer Verlag.

Stanley, R. P. (1997). *Enumerative Combinatorics - Volume I*. Number 49 in Cambridge Studies in Advance Mathematics. Cambridge University Press.

Stewart, G. W. (1998). *Matrix Algorithms: Basic Decompositions*, Volume I. Society for Industrial and Applied Math.

Sturmfels, B. (1996). *Gröbner Bases and Convex Polytopes*. American Mathematical Society.

Takemura, A. and S. Aoki (2004). Some characterizations of minimal markov basis for sampling from discrete conditional distributions. *The Annals of Statistics 56*, 1–17.

Tarjan, R. E. (1985). Decomposition by clique separators. *Discrete Mathematics 55*, 221–232.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley Series in Probability & Statistics. Wiley.

Ziegler, M. G. (1998). *Lectures on Polytopes*. New York: Springer-Verlag.