On Maximum Likelihood Estimation in Log-Linear Models

Alessandro Rinaldo*
Department of Statistics
Carnegie Mellon University

Abstract

In this article, we combine results from the theory of linear exponential families, polyhedral geometry and algebraic geometry to provide analytic and geometric characterizations of log-linear models and maximum likelihood estimation. Geometric and combinatorial conditions for the existence of the Maximum Likelihood Estimate (MLE) of the cell mean vector of a contingency table are given for general log-linear models under conditional Poisson sampling. It is shown that any log-linear model can be generalized to an extended exponential family of distributions parametrized, in a mean value sense, by points of a polyhedron. Such a parametrization is continuous and, with respect to this extended family, the MLE always exists and is unique. In addition, the set of cell mean vectors form a subset of a toric variety consisting of non-negative points satisfying a certain system of polynomial equations. These results of are theoretical and practical importance for estimation and model selection.

1 Introduction

Log-Linear models are a powerful statistical tool for the analysis of categorical data with applications in a variety of scientific areas, ranging from social and biological sciences, to medicine, disclosure limitation problems, data-mining, image analysis, finger-printing, language processing and genetics. In the last years, their importance and usage have increased greatly with the compilation and diffusion of large databases in the form of sparse contingency tables, where most of the cell entries are very small or zero counts. In these instances, the the Maximum Likelihood Estimate (MLE) of the expected value of the cell mean vector, fundamental for assessment of fit, model selection and interpretation, is very likely to be undefined, or nonexistent.

In log-linear modeling, the existence of the MLE is essential for the usual derivation of large sample χ^2 approximations to numerous measures of fit (Bishop et al., 1975; Agresti, 2002; Cressie and Read, 1988) which are utilized to perform hypothesis testing and model selection. If the distribution of the goodness of fit statistics 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 estimate in order to quantify the discrepancy of the the observed data from the fitted values. In addition, the existence of the MLE is required for obtaining a limiting distribution in the double-asymptotic approximations of the likelihood ratio and Pearson's χ^2 statistic for tables in which both the sample size and the number of cells are allowed to grow unbounded, a setting studied, in particular, by Morris (1975), Haberman (1977)

*Email: arinaldo@stat.cmu.edu

and Koehler (1986) (see Cressie and Read, 1988, for a complete literature review). If the MLE is not defined, the inferential procedures mentioned above may not be applicable or, at a minimum, require alteration.

The problem of nonexistence of the MLE has long been known to relate to the presence of zero cell counts in the observed table (see, in particular, Haberman, 1974; Bishop et al., 1975). Sampling zeros may be thought of as missing bits of information. When they occur in specific patterns inside the table, such as those in the tables of Section 3.1, 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 the MLE does not exist. 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) and discussed in Table **a)** of Section 3.1. 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 lead directly to implementable numerical procedures and also fails to suggest alternative methods of inference for the case of an undefined MLE. Despite these deficiencies, Haberman (1974)'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 and the only indication of nonexistence is a lack of convergence of whatever algorithm is used to compute the MLE. 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, in particular, the examples in Fienberg and Rinaldo (2006b) and Rinaldo (2005, Chapter 1). 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.

In this article, we we propose a general framework for log-linear model analysis and we derive analytic and geometric properties of the maximum likelihood estimation for log-linear models. Motivated by the the recent advances in the field of algebraic statistics (Diaconis and Sturmfels, 1998; Pistone et al., 2000; Pachter and Sturmfels, 2005), throughout this article, we demonstrate and then take advantage of some connections between algebraic and polyhedral geometry and the theory of exponential families. First, we derive novel geometric and combinatorial conditions for the existence of the MLE for a large class of log-linear models that generalize results currently available in the statistical literature and that are suited to numerical implementation. We then show that log-linear models can be associated with extended linear exponential families of distributions parametrized, in a mean value sense, by non-negative points lying on toric varieties. Within the framework of extended exponential families, the MLE, which we call extended MLE, always exists and is unique. We then derive various analytical and geometric properties of the extended MLE and discuss their implications. Our results build upon important contributions of many authors. In particular, we mention Haberman (1974), Bardorff-Nielsen (1978), Brown (1986), Fienberg et al. (1980), Lauritzen (1996), Diaconis and Sturmfels (1998), Geiger et al. (2006), Sturmfels (2003), Csiszár and Matúš (2001, 2003, 2005) and Eriksson et al. (2005).

The article is organized as follows. In Section 2 we introduce linear exponential families for

discrete distributions over finite sets. We show how this hypothesis leads naturally to the study of contingency tables and the formulation of a log-linear representation on the cell mean vector. We consider sampling schemes specified by linear constraints on the cell counts and determine their effects on the estimability of the parameters of interest. Section 3 provides general results for existence of the MLE for log-linear models using the theory of exponential families and basic results from polyhedral geometry. We show that the existence of the MLE is equivalent for the vector of the observed sufficient statistics to belong to the relative interior of a polyhedron determined by the underlying log-linear model parameters and the sampling constraints. Section 4 defines extended exponential families for the sufficient statistics and cell counts. The construction proceeds through various steps. First, we show that maximizing the log-likelihood function is a well defined problems which, under mean value parametrization, has always a limiting solution. Then, we show how to take advantage of the geometric properties of the convex support to define the extended exponential families for the sufficient statistics and we describe their properties. Finally, we derive an extended exponential family of distribution for the cell counts and prove that it can be conveniently parametrized by a set of points homeomorphic to a polyhedral cone. These points are called the extended MLEs and, for Poisson and product-multinomial scheme, corresponds to the sequential closure of the set of all cell mean vectors. In Section 5 we show that the set of all extended MLEs can be represented as a toric variety and we give a geometric characterization of the parameter space for log-linear models, under mean value parametrization.

We conclude this introduction by describing the notation used throughout the article. 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). 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 with $(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{I}} x_i y_i$, 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 as $i_s =< \mathbf{i}_s >$. The set of vectors in $\mathbb{R}^{\mathcal{I}}$ with non-negative coordinates will be denoted with $\mathbb{R}^{\mathcal{I}}_{\geq 0}$ and the support $\operatorname{supp}(\mathbf{x})$ of a vector $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ is the set $\{i \in \mathcal{I} \colon x_i \neq 0\}$. Functions and relations on vectors will be taken component-wise, unless otherwise specified. For example, for $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, $\exp^{\mathbf{x}} = \{\exp^{x_i} \colon i \in \mathcal{I}\}$. The cardinality of a numerable set B will be denoted by $|B|$.

2 Exponential Families, Contingency Tables and Sampling Schemes

Log-linear model analysis is concerned with the study and characterization of the joint distribution of the K categorical variables $(X_1, \ldots, X_K)$. This distribution is assumed to belong to the exponential family of probabilities $\{P_{\boldsymbol{\eta}}\}$ on $\mathcal{I}$ with density with respect to the counting measure of the form

$$p_{\boldsymbol{\eta}}(i) = P_{\boldsymbol{\eta}}(\{i\}) = \exp^{(\boldsymbol{\eta}, T(i)) - \phi(\boldsymbol{\eta})}, \tag{2}$$

where $T : \mathcal{I} \mapsto \mathbb{Z}^d \setminus \{\mathbf{0}\}$ is given by $T(i) = \mathbf{a}_i$, ϕ is a read-valued function from $\mathbb{R}^d$ into $\mathbb{R}$ which serves as a normalizing constant and $\boldsymbol{\eta}$ belongs to the natural parameter space $\mathbf{H} = \{\boldsymbol{\eta} \in \mathbb{R}^d : \exp\{\phi(\boldsymbol{\eta})\} < \infty\}$ (see Diaconis and Sturmfels, 1998; Geiger et al., 2006).

Data are collected by observing N independent realizations of the K variables and typically take the form of an unordered random sequence of label combinations $(L_1, \ldots, L_N)$, with $L_j \in \mathcal{I}$ for each $j = 1, \ldots, N$, where N too can be random. A contingency table $\mathbf{n}$ is a (non-minimal) sufficient statistic for $\boldsymbol{\eta}$ obtained by counting the number of times each cell has appeared in the sample. Formally, a contingency table is a random function $\mathbf{n} \in \mathbb{R}^{\mathcal{I}}$ given by $\mathbf{n}(i) = |\{j : L_j = i\}|$. A minimal sufficient statistic for $\boldsymbol{\eta}$ is instead $\mathbf{t} = \sum_{j=1}^{N} T(L_j) = \sum_{i \in \mathcal{I}} \mathbf{n}(i)\mathbf{a}_i = \mathrm{A}\mathbf{n}$, where A is the $d \times |\mathcal{I}|$ design matrix whose i-th column is the vector $\mathbf{a}_i$. Inference on $\boldsymbol{\eta}$ is performed, more conveniently, by studying the distribution of the random vector of counts $\mathbf{n}$. Specifically, given a σ-finite measure ν on $\mathbb{N}^{\mathcal{I}}$, defined below, the distribution of $\mathbf{n}$ belongs to the standard exponential family of probability distributions $\{\mathbb{P}_{\boldsymbol{\eta}}\}_{\eta \in \mathbf{E}}$ generated by A and ν with ν-density

$$p_{\boldsymbol{\eta}}(\mathbf{x}) = \exp\{(\mathbf{t}, \boldsymbol{\eta}) - \psi(\boldsymbol{\eta})\}, \tag{3}$$

where $\mathbf{t} = \mathrm{A}\mathbf{x}$ is the sufficient statistic, $\psi(\boldsymbol{\eta}) = \log \int \mathrm{e}^{(\boldsymbol{\eta}, \mathbf{t})} \nu(d\mathbf{x})$ is the function and $\mathbf{H} = \{\boldsymbol{\eta} : \psi(\boldsymbol{\eta}) < \infty\} \subseteq \mathbb{R}^d$ the natural parameter space.

The log-linear modeling framework is based on the representation of the cell mean vector $\mathbf{m} = \mathbb{E}[\mathbf{n}]$ by means of the linear subspace $\mathcal{M}$ of $\mathbb{R}^{\mathcal{I}}$, called *log-linear subspace*, spanned by the rows of the design matrix A. Specifically, $\mathcal{M}$ consists of all the log cell mean vectors $\boldsymbol{\mu} = \log(\mathbf{m})$. By Proposition 2.1 below, this is in fact equivalent to assuming the family of distribution (2) over $\mathcal{I}$.

The distributions of the cell counts and the quality of the inference depend on $\mathcal{M}$ and also on the type of sampling scheme utilized in the collection of the data. This work considers only sampling designs specified by linear constraints, requiring each observed table $\mathbf{n}$ to satisfy linear forms. Formally, let $\mathcal{N} \subsetneq \mathcal{M}$ be a linear subspace and denote by $\mathrm{P}_{\mathcal{N}}$ the projection matrix onto $\mathcal{N}$. Then, $\mathcal{N}$ specifies sampling restrictions of the form $\mathbf{c} = \mathrm{P}_{\mathcal{N}}\mathbf{n}$ for a constant vector $\mathbf{c}$. Equivalently, if $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)$ are m vectors spanning $\mathcal{N}$, the sampling constraints are $(\boldsymbol{\gamma}_j, \mathbf{n}) = c_j$, for constants c_j, $j \leq 1 \leq m$. Let $S(\mathcal{N}) = \{\mathbf{x} \in \mathbb{N}^{\mathcal{I}} : \mathrm{P}_{\mathcal{N}}\mathbf{x} = \mathbf{c}\}$ denote the set of all possible tables compatible with the constraints determined by $\mathcal{N}$ and assume that $S(\mathcal{N}) \neq \emptyset$. Note that, if $\mathcal{N} = \{\mathbf{0}\}$, then the sampling is unconstrained. Given a constraint subspace $\mathcal{N}$, the base measure $\nu_{\mathcal{N}}$ for the exponential family (3) is defined as

$$\nu_{\mathcal{N}}(\mathbf{x}) = \begin{cases} \nu(\mathbf{x}) := \frac{1}{\prod_i \mathbf{x}(i)!} \mathbf{1}_{\mathbf{x} \in \mathbb{N}_+^{\mathcal{I}}} & \text{if } \mathcal{N} \\ \text{restriction of } \nu \text{ on } S(\mathcal{N}) & \text{otherwise,} \end{cases} \tag{4}$$

where, $\mathbf{1}_{\mathbf{x} \in B}$ is the indicator function of the set B. Note in particular that, for any subspace $\mathcal{N} \subseteq \mathbb{R}^{\mathcal{I}}$, $\nu_{\mathcal{N}}(\mathbb{R}^{\mathcal{I}}) \leq \mathrm{e}^{|\mathcal{I}|} < \infty$. Letting $\mathcal{M} \ominus \mathcal{N} = \mathcal{M} \cap \mathcal{N}^{\perp}$, we will be making the following assumption:

Sampling assumption (S): there does not exist any vector $\boldsymbol{\gamma} \in \mathcal{M} \ominus \mathcal{N}$, such that $(\boldsymbol{\gamma}, \mathbf{n}) = c$ a.e.-$\nu_{\mathcal{N}}$, for any $c \in \mathbb{R}$.

Assumption **(S)** guarantees that the constraint subspace $\mathcal{N}$ encodes all possible affine dependencies among the observable tables.

Sampling schemes of this type are called ***conditional Poisson sampling schemes*** (Haberman, 1974) because they induce a Poisson conditional distributions of the counts $\mathbf{n}$ given $S(\mathcal{N})$. A general expression for the conditional Poisson distribution is

$$p_{\boldsymbol{\eta}}(\mathbf{x}) d\nu_{\mathcal{N}}(\mathbf{x}) = \frac{\frac{1}{\prod_i \mathbf{x}(i)!} \exp\{(\mathbf{t}, \boldsymbol{\eta})\}}{\sum_{\mathbf{x}' \in S(\mathcal{N})} \frac{1}{\prod_i \mathbf{x}'(i)!} \exp\{(\mathbf{t}', \boldsymbol{\eta})\}} \frac{d\nu_{\mathcal{N}}}{d\nu}(\mathbf{x}), \tag{5}$$

where $\mathbf{t}' = \mathrm{A}\mathbf{x}'$ and $\frac{d\nu_{\mathcal{N}}}{d\nu}(\mathbf{x}) = \mathbf{1}_{\mathbf{x} \in S(\mathcal{N})}$. Equation (5) defines an exponential family of distributions with log-partition function $\psi_{\nu_{\mathcal{N}}}$ given by the denominator of the right hand side expression in (5) and corresponding parameter space

$$\mathbf{H}_{\nu_{\mathcal{N}}} = \{\boldsymbol{\eta} : \psi(\boldsymbol{\eta}) < \infty\} \subseteq \mathbb{R}^d\}.$$

Note that the probability mass functions for these conditional distributions, and their moments, typically do not have closed forms. The most common sampling schemes, which happen to posses densities in closed form, are:

- **Poisson scheme**
  $\mathcal{N} = \{\mathbf{0}\}$. There are no restrictions on $\mathbf{n}$.

- **Multinomial sampling**
  $\mathcal{N}$ is the set of constant functions on $\mathcal{I}$. There is only one linear restriction on $\mathbf{n}$ of the form $(\mathbf{1}_{\mathcal{I}}, \mathbf{n}) = N$, where $\mathbf{1}_{\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ is the vector of ones, and the grand total is a positive integer N fixed by design. The conditional distribution of the counts $\mathbf{n}$ given the constraints is multinomial with size N.

- **Product-multinomial sampling**
  Let $\mathcal{B}_1, \ldots, \mathcal{B}_r$ be a partition of $\mathcal{I}$. Under the product-multinomial sampling, the conditional distribution of the cell counts $\mathbf{n}$ is the product of independent multinomials of sizes N_j, $j = 1, \ldots, r$, each supported on the corresponding class $\mathcal{B}_j$. Formally, let $\boldsymbol{\chi}_j$ be the indicator function of $\mathcal{B}_j$ given by

  $$\boldsymbol{\chi}_j(i) = \begin{cases} 1 & \text{if } \ i \in \mathcal{B}_j \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

  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 are $(\mathbf{n}, \boldsymbol{\chi}_j) = N_j$, for integer constants N_j. The spanning vectors of $\mathcal{N}$ are often defined in a simpler way. 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 $\mathcal{B}_j = \{i \in \mathcal{I} \colon i_b = j\}$. Then, the sets $\mathcal{B}_j$ form a partition of $\mathcal{I}$, and $\mathcal{N}$ is the r-dimensional subspace spanned by the vectors $\{\boldsymbol{\chi}_j\}_{j \in \mathcal{I}_b}$, where $\boldsymbol{\chi}_j$ is defined as in (6) and $r = |\mathcal{I}_b|$. Some authors, such as Lauritzen (1996), use this partitions to define product-multinomial scheme. The multinomial scheme is a special case of product-multinomial schemes, corresponding to the trivial one-class partition of $\mathcal{I}$ with indicator function $\mathbf{1}_{\mathcal{I}}$.

The following results shows that assuming the linear exponential family representation (2) for the joint distribution of the variables is equivalent to specifying a log-linear models for the cell counts $\mathbf{n}$.

Proposition 2.1. *Let $\boldsymbol{\mu} = \log \mathbf{m}$, where $\mathbf{m}$ is the unconditional cell mean vector corresponding to Poisson sampling. Then,*

i. Model (2) holds if and only if $\boldsymbol{\mu} \in \mathcal{M}$.

ii. The conditional distribution of the table $\mathbf{n}$ given the sampling constraints belongs to a linear exponential family parametrized by $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu}$ and for which $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\mathbf{n}$ is a minimal sufficient statistic.

If $\mathcal{M} = \mathcal{N}$, the constraints are so restrictive that no inference is possible because the sufficient statistics are constant functions over $S(\mathcal{N})$. This situation is characteristic of generalized hypergeometric types of distributions, which can be seen as special cases of conditional Poisson sampling. Those cases are uninteresting from the point of view of maximum likelihood estimation, as the distribution of the counts does not depend on the parameters of interest.

We conclude this section by pointing out that the sampling constraints may be chosen to be so restrictive that no tables with all positive entries can be observed. In this case, the sampling scheme is said to be improper. Formally,

Definition 2.2. A sampling scheme defined by the subspace $\mathcal{N} \subsetneq \mathcal{M}$ is called *proper* if there is no coordinate $i \in \mathcal{I}$ such that $\mathbf{n}(i) = 0$ a.e.-$\nu_{\mathcal{N}}$ and *improper* otherwise.

Under improper schemes, some cells have zero probability of being observed only because of sampling constraints. The notion of improper sampling scheme is completely different than the one of structural zeros, which are in fact independent of the sampling scheme adopted. We assume here that there are no structural zeros, namely each cell has a strictly positive probability of being observed, prior to imposing sampling limitations. Indeed, this assumption was implicitly used at the beginning of this section. Using Lemma 4.6 we will show how to use improper sampling to formalize a reduced information content in the sufficient statistics and, in particular, to identify the set of cell mean counts for which the MLE cannot be computed. For the remainder of this article, we will always assume proper sampling, unless otherwise stated.

3 Exponential Families for Count Data and MLE

The study of log-linear models and the conditions for the existence of the MLE can be cast inside the more general framework of the theory of standard exponential families (see, in particular, Bardorff-Nielsen, 1978; Brown, 1986).

Consider the exponential family of distribution (3) and let $\mu_{\mathcal{N}}$ be the finite measure on $\mathbb{Z}^d$ induced by $\nu_{\mathcal{N}}$ and the linear transformation determined by the matrix A.

Definition 3.1. The *convex support* $\mathcal{C}_{\mathcal{N}}$ associated with $\nu_{\mathcal{N}}$ and A is the closure of the convex hull of the points in the support of the induced measure $\mu_{\mathcal{N}}$.

Csiszár and Matúš (2001, 2005) prove that it is more convenient to examine instead the *convex core* of the induced measure $\mu_{\mathcal{N}}$, defined as the intersection of all convex closed sets of full $\mu_{\mathcal{N}}$-measure. In general, the convex core can be smaller than the convex support, although they have

the same relative interior and closure. However, for the present settings, convex core and convex support coincide, a fact of immediate verification.

Definition 3.2. The mapping $\Lambda : \mathbf{H}_{\nu_\mathcal{N}} \to \mathcal{C}_\mathcal{N}$ given by $\Lambda(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}(\mathrm{A}\mathbf{x}) = \nabla\psi(\boldsymbol{\eta})$ is called the mean-value parametrization, where ∇ denotes the gradient.

Under minimality and regularity, Λ defines a homeomorphism between the interior of the natural parameter space and the relative interior $\mathrm{ri}(\mathcal{C}_\mathcal{N})$ of the convex support. When the exponential family is not expressed in minimal form, the mean value parametrization is no longer a homeomorphism but it is a surjective map over $\mathrm{ri}(\mathcal{C}_\mathcal{N})$. This means that distributions belonging to the same linear exponential family are different if and only if they have different means.

Definition 3.3. Let $\ell(\boldsymbol{\eta}, \mathbf{x}) = \log p_{\boldsymbol{\eta}}(\mathbf{x})$ and $\widehat{\boldsymbol{\eta}}(\mathbf{x}) = \{\boldsymbol{\eta} \in \mathbf{H}_{\nu_\mathcal{N}} : \ell(\boldsymbol{\eta}, \mathbf{x}) = \sup_{\boldsymbol{\eta} \in \mathbf{H}_{\nu_\mathcal{N}}} \ell(\boldsymbol{\eta}, \mathbf{x})\}$. Any point $\widehat{\boldsymbol{\eta}} \in \widehat{\boldsymbol{\eta}}(\mathbf{t})$ is called a maximum likelihood estimate of $\boldsymbol{\eta}$. If $\widehat{\boldsymbol{\eta}}(\mathbf{x}) = \emptyset$, then we say that the MLE does not exist.

A fundamental result for the existence of the MLE for minimal, full and regular exponential families is stated in the next theorem.

Theorem 3.4 (Brown, 1986, Theorem 5.5)**.** *The MLE $\widehat{\boldsymbol{\eta}}$ exists and is unique if and only if $\mathbf{t} \in \mathrm{ri}(\mathcal{C}_\mathcal{N})$ and, if existent, it satisfies the moment equation $\Lambda(\widehat{\boldsymbol{\eta}}) = \widehat{\mathbb{E}(\mathrm{A}\mathbf{n})} = \mathbf{t} \in \mathrm{ri}(\mathcal{C}_\mathcal{N})$.*

Since the exponential family (5) for the cell counts is typically neither minimal nor full, we choose to study instead the distribution of the minimal sufficient statistic. Let A^* be a $k \times |\mathcal{I}|$ full row-rank integer matrix whose rows span $\mathcal{M} \ominus \mathcal{N}$, where $k = \dim(\mathcal{M} \ominus \mathcal{N})$. Then, $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu} = (\mathrm{A}^*)^\top\boldsymbol{\theta}$, for some natural parameter $\boldsymbol{\theta} \in \mathbb{R}^k$. Note also that, by the sampling assumption **(S)**, the linear map from $\mathbb{R}^\mathcal{I}$ into $\mathbb{R}^k$ specified by A^* does not induce any affine dependencies on its image.

The following theorem provides conditions for the existence of the MLE for the natural parameter of the family (5). It exploits Theorem 3.4 and the geometric properties of the convex supports.

Theorem 3.5. *The conditional Poisson model with constraint subspace $\mathcal{N}$ induces for $\mathbf{z} = \mathrm{A}^*\mathbf{n}$ a minimal, regular and full linear exponential family of order $k = \dim(\mathcal{M} \ominus \mathcal{N})$ and natural parameter space $\mathbb{R}^k$. The MLE of $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu}$ exists and is unique if and only if $\mathbf{t} = \mathrm{A}\mathbf{n}$ belongs to the relative interior of the k-dimensional polyhedron $\mathrm{P}_\mathrm{A} = \{\mathrm{A}\mathbf{x} : \mathbf{x} \geq \mathbf{0}, (\boldsymbol{\gamma}_j, \mathbf{x}) = c_j, j = 1, \ldots, m\}$.*

Equation (26) in the proof of Theorem 3.5 gives a more refined result. Since every polyhedron is in fact the Minkowski sum of a polytope and a polyhedral cone (see, for example, Theorem 1.2 in Ziegler, 1998), we can conclude that the convex support $\mathcal{C}_\mathcal{N}$ for the sufficient statistic of a conditional Poisson sampling scheme is a polyhedron which can be obtained as the coordinate projection of the Minkowski sum of a polyhedral cone and a polytope. In particular, the polytope component arises from the linear forms defining the sampling constraints.

Under Poisson sampling there are no constraints, hence the convex support is the polyhedral cone $\mathrm{C}_\mathrm{A} = \mathrm{cone}(\mathrm{A})$ generated by the columns of the A, called the *marginal cone* (Eriksson et al., 2005). For multinomial sampling the convex support is instead a polytope, whose homogeneization is precisely C_A. As a result, the two polyhedra are combinatorially equivalent. Under product-multinomial sampling, the convex support is a polytope whose dimension is smaller than the one arising from the multinomial scheme. For general sampling schemes, the combinatorial equivalence with C_A is not preserved because typically the cone C_A has more faces than any other polyhedron

combinatorially equivalent to the convex support $\mathcal{C}_{\mathcal{N}}$, as the next example shows (see also Lemma 3.9 below). In general, a sufficient condition for $\mathcal{C}_{\mathcal{N}}$ to be bounded (hence a polytope) is that $\mathcal{N}$ contains vectors of the same sign, as it is the case with standard hierarchical models.

Example 3.6. Consider the simple case of a 2^2 table under the model of independence, for which the matrix A can be taken to be

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}. \tag{7}$$

The cell labels set $\mathcal{I} = \{(11), (12), (21), (22)\}$ is identified with the list $\{1, 2, 3, 4\}$ determined by Equation (1). Under both Poisson and multinomial sampling, the marginal cone A has 4 facets, with vertex-facet incidence vectors corresponding precisely to the rows of A. If we use instead product-multinomial sampling with constraint subspace spanned by the columns of the matrix

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

then only the two facets corresponding to the last two rows of A can be observed. ■

It is worth to point out that mixed sampling schemes such as the Poisson-multinomial schemes proposed by Lang (2004), can be naturally accommodated within this framework. In fact, any sampling scheme requiring the cell counts to satisfy linear constraints will produce a polyhedral representation of the convex support.

As an illustration of the exponential family approach, a novel proof of the well known result on the equivalence of the MLE for the mean vector under Poisson and product-multinomial schemes (see, in particular, Birch, 1963; Haberman, 1974) is given in the following theorem.

Theorem 3.7. *Provided that $\mathcal{N} \subset \mathcal{M}$, the MLE $\widehat{\mathbf{m}}$ of the cell mean vector under Poisson sampling scheme exists if and only if the MLE of the conditional cell mean vector under product-multinomial sampling schemes exists. In this case, they coincide, are unique and satisfy the moment equations $\mathcal{P}_{\mathcal{M}}\widehat{\mathbf{m}} = \mathcal{P}_{\mathcal{M}}\mathbf{n}$.*

Note that the equivalence is not guaranteed if the condition $\mathcal{N} \subset \mathcal{M}$ fails (see also Lang, 1996). The previous result implies that the sampling constraints for product-multinomial schemes are mild, in the sense that the MLE of the conditional mean vector is identical to the unconditional one. In general, this is not the case under general conditional Poisson schemes, because, although the MLE of the conditional cell mean vector $\widehat{\mathbf{m}}$ satisfies the moment equations, it does not necessarily satisfy $\log \widehat{\mathbf{m}} \in \mathcal{M}$.

The density of the minimal sufficient statistics $\mathbf{z} = \mathrm{A}^*\mathbf{n}$ with respect to $\mu_{\mathcal{N}}$ is

$$p_{\boldsymbol{\theta}}(\mathbf{z}) = \exp\{(\mathbf{z}, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\} \qquad \boldsymbol{\theta} \in \Theta, \tag{8}$$

where $\psi(\boldsymbol{\theta}) = \int_{\mathbb{R}^k} \exp^{(\boldsymbol{\theta}, \mathbf{z})} d\mu_{\mathcal{N}}(\mathbf{z})$ and $\Theta = \{\boldsymbol{\theta} : \ \psi(\boldsymbol{\theta}) < \infty\} = \mathbb{R}^k$. We will be denoting the convex support for this family with the same symbol $\mathcal{C}_{\mathcal{N}}$, although it is now clear that this polyhedron is full-dimensional, A^* being of full-row rank. The densities in (8) and (5) are related in the following way.

Lemma 3.8. *Assume, without loss of generality that* A *is of full-row rank. For all* $\mathbf{n} \in S(\mathcal{N})$ *and for any* $\boldsymbol{\theta} \in \mathbb{R}^k$,

$$p_{\boldsymbol{\theta}}(\mathbf{z}) = p_{\boldsymbol{\eta}}(\mathbf{n}), \quad \forall \boldsymbol{\eta} \in \mathbf{H}_{\nu_{\mathcal{N}}} \text{ such that } \mathrm{H}\boldsymbol{\eta} = \boldsymbol{\theta}, \tag{9}$$

where H *is a matrix independent of* $\boldsymbol{\eta}$.

Lemma 3.8 shows that imposing sampling constraint results in lack of parameter identifiability. However we will see at the end of Section 4.2 that it is always possible to resolve this non-identifiability, i.e. H defines in fact a bijection on $\mathbf{H}_{\nu_{\mathcal{N}}}$, so that the MLE of $\mathrm{P}_{\mathcal{M}\oplus\mathcal{N}}\boldsymbol{\mu}$ always identifies one vector in $\mathcal{M}$.

There is a correspondence between the convex support $\mathcal{C}_{\mathcal{N}}$ and the marginal cone C_{A}. In fact, $\mathcal{C}_{\mathcal{N}}$ is isomorphic to the polyhedron resulting from the intersection of C_{A} with the hyperplane defined by the sampling constraints. More formally,

Lemma 3.9. *There exists a linear injection* f_{A} *of* $\mathcal{C}_{\mathcal{N}}$ *into* C_{A} *such that each face of* $\mathcal{C}_{\mathcal{N}}$ *is mapped into a face of* C_{A}.

The map f_{A} is never a bijection unless $\mathcal{N}$ is the trivial subspace $\{\mathbf{0}\}$. In this case, A^* and A have the same rank and the corresponding marginal cones C_{A} and $\mathrm{C}_{\mathrm{A}^*}$ are isomorphic. Using the previous lemma, a general condition for the existence of the MLE of $\mathcal{P}_{\mathcal{M}\oplus\mathcal{N}}\boldsymbol{\mu}$ and, hence, of the natural parameter $\boldsymbol{\theta}$, can be established using the marginal cone C_{A}.

Corollary 3.10. *Under proper sampling, the MLE of* $\mathcal{P}_{\mathcal{M}\oplus\mathcal{N}}\boldsymbol{\mu}$ *exists and is unique if and only if* $\mathbf{t} \in \mathrm{ri}(\mathrm{C}_{\mathrm{A}})$.

This result generalizes Theorem 2.2 and 2.5 in Haberman (1974) (see Appendix B) and Corollary 3 in Eriksson et al. (2005).

3.1 Examples

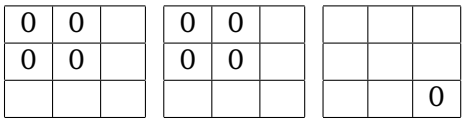
The polyhedral characterization of the conditions for the existence of the MLE permits to generate novel examples of patterns of sampling zeros causing non-existence of the MLE without producing null margins. The examples presented below were obtained using `polymake` (Gawrilow, 2000), a software for the algorithmic treatment of convex polyhedra. These examples suggest that the combinatorial complexity of hierarchical log-linear models can be quite significant. Below we denote the generating class of a hierarchical log-linear model on K variables as represented as a class of subsets of $\{1, \ldots, K\}$.

a) 2^3 table and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$ of no-second-order interaction (Haberman, 1974). The MLE is not defined because the pattern zeros exposes one of the 16 facets of the marginal cone. See Table 1 in Eriksson et al. (2005) and Section 5 in Fienberg and Rinaldo (2006b) for a more general result concerning binary variables and log-linear models of no-$(K-1)$st interaction. This example has been for a long time the only published instance of "pathological" (Bishop et al., 1975, page 115) tables with positive margins and non-existent MLE.

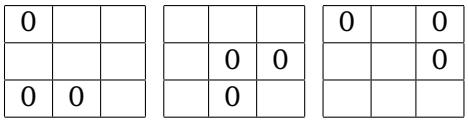
0	

	0

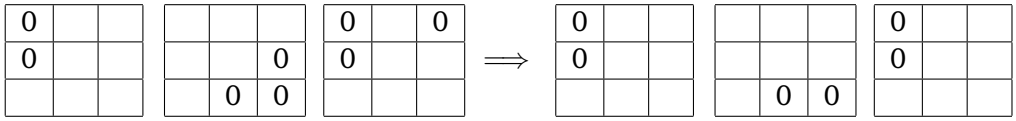
b) 3^3 table and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$. The MLE is not defined because the pattern of zeros exposes one of the 207 facets of the marginal cone.



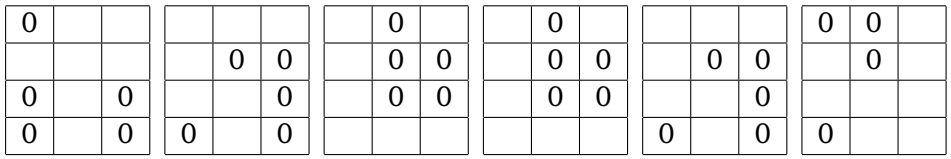
Another facet of the same marginal cone is given by



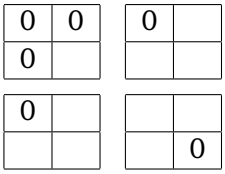
In this third and final example, two sampling zeros in the left table are not reported in the right table because they don't affect the existence of the MLE and, in fact, correspond to positive cell mean values for the extended MLE (see later Section 4). The table on the right exposes a facet of the marginal cone.



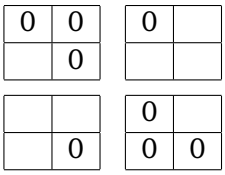
c) $4 \times 3 \times 6$ table and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$. The MLE is not defined because the pattern of zeros exposes one of 153,858 facets of the marginal cone.



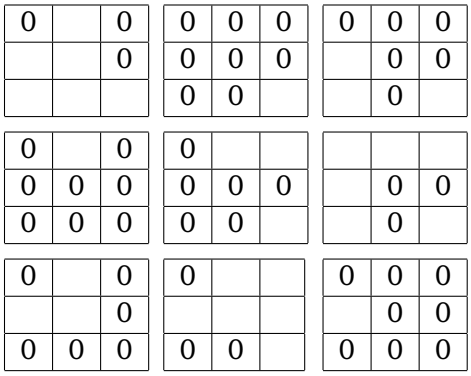
d) 2^4 table and the non-graphical model $\{\{1,2\},\{2,3\},\{3,4\},\{1,4\},\{1,3\},\{2,3\}\}$. The MLE is not defined because the pattern of zeros exposes one of the 56 facets of the marginal cone.



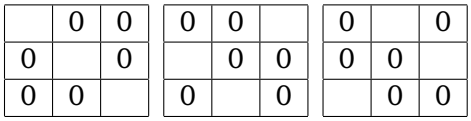
e) 2^4 table and the 4-cycle model $\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$. The MLE is not defined because the pattern of zeros exposes one of the 24 facets of the marginal cone.



f) 3^4 table and the 4-cycle model $\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$. The MLE is not defined because the pattern of zeros exposes one of the 1,116 facets of the marginal cone.



g) 3^3 table and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$ (Fienberg and Rinaldo, 2006b). The MLE is defined, despite the table being very sparse, because no facet of the marginal cone is exposed.



4 Extended Exponential Families and the Extended MLE

This section defines extended linear exponential families for discrete data and describes some of their properties. The construction builds on results by Bardorff-Nielsen (1978), Brown (1986) and Csiszár and Matúš (2003, 2005).

4.1 Extended Exponential Families for Sufficient Statistics

Consider the minimal, regular and full exponential families (8) and recall that the natural parameter space Θ is $\mathbb{R}^k$ and the convex support $\mathcal{C}_{\mathcal{N}}$ is a polyhedron defined by the design matrix and the sampling linear constraints implied by the subspace $\mathcal{N} \subsetneq \mathcal{M}$.

We first show that the supremum of the log-likelihood function is always finite and attainable. In fact, as noted by Haberman (1974), the non-existence of the MLE does not imply that the log-likelihood function explodes. In fact, it only implies that the supremum is realized in the limit by sequences of points in the natural parameter space with exploding norm. The same result, in less generality, was also proved by Lauritzen (1996, Section 4.2.3). To this extent, consider the *sup-log-likelihood function*, introduced by Bardorff-Nielsen (1978).

Definition 4.1. The map $\psi^* : \mathbb{R}^k \to \mathbb{R}$ given by $\psi^*(\boldsymbol{\xi}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^k} \ln p_{\boldsymbol{\theta}}(\boldsymbol{\xi})$ is called the sup-log-likelihood function.

Let $\mathrm{dom}(\psi^*) = \{\boldsymbol{\xi} \in \mathbb{R}^k : \psi^*(\boldsymbol{\xi}) < \infty\}$ denote the effective domain of ψ^*. Note that the sup-log-likelihood function, being the conjugate function of the function ψ, is defined on the whole $\mathbb{R}^k$ rather than just on $\mathrm{supp}(\mu_{\mathcal{N}})$. Although the sufficient statistic $\mathbf{z}$ may lie on the boundary of the convex support (hence causing the MLE not to be defined), the supremum of the log-likelihood is always finite, i.e. $\mathrm{dom}(\psi^*) \subset \mathcal{C}_{\mathcal{N}}$. In fact, the last inclusion is an identity.

Theorem 4.2. *The sup-log-likelihood function ψ^* is a closed, essentially smooth strictly convex function such that $\mathrm{dom}(\psi^*) = \mathcal{C}_{\mathcal{N}}$.*

The proof of the last result, given in the appendix, relies on standard results in convex analysis (see, in particular, Rockafellar, 1970).

Remark. From Equation (28) in the proof of Theorem 4.2, a well known fact can be derived (see, for example, Jordan and Wainwright, 2003), linking maximum likelihood estimation and the Bolzmann-Shannon entropy, given, for $\boldsymbol{\theta} \in \Theta$, by

$$H(\boldsymbol{\theta}) = -\int \left(\log p_{\boldsymbol{\theta}}(\mathbf{z})\right) p_{\boldsymbol{\theta}}(\mathbf{z}) d\mu_{\mathcal{N}}(\mathbf{z}).$$

Corollary 4.3. *If $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$, $\psi^*(\boldsymbol{\xi}) = -H(\Lambda^{-1}(\boldsymbol{\xi}))$. If $\boldsymbol{\xi} \in \mathrm{bd}(\mathcal{C}_{\mathcal{N}})$, then for any sequence $\{\boldsymbol{\xi}_i\} \subset \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ with $\lim_i \boldsymbol{\xi}_i = \boldsymbol{\xi}$, $\psi^*(\boldsymbol{\xi}) = \lim_i -H(\Lambda^{-1}(\boldsymbol{\xi}_i))$. If $\boldsymbol{\xi} \notin \mathcal{C}_{\mathcal{N}}$, $\psi^*(\boldsymbol{\xi}) = \infty$.*

The previous result is used in Grunwald and Dawid (2004), where connections between maximum likelihood estimation, information theory and minimaxity are explored.

Let F be any proper face of $\mathcal{C}_{\mathcal{N}}$ and μ_F be the restriction of $\mu_{\mathcal{N}}$ on F. Associate to F a new linear exponential family of distributions having base measure μ_F and convex support F, with log-partition function $\psi_F(\boldsymbol{\theta}) = \log \int \mathrm{e}^{(\boldsymbol{\theta},\mathbf{x})} d\mu_F(\mathbf{x})$, parameter space $\Theta_F = \{\boldsymbol{\theta} \in \mathbb{R}^k \colon \psi_F(\boldsymbol{\theta}) < \infty\}$ and densities

$$p_{\boldsymbol{\theta}}(\mathbf{z}) = \exp\{(\mathbf{z}, \boldsymbol{\theta}) - \psi_F(\boldsymbol{\theta})\}, \quad \mathbf{z} \in F.$$

This new family is no longer minimal because $\dim(F) < \dim(\mathcal{C}_{\mathcal{N}})$ but it is still full and regular because $\mathbb{R}^k = \Theta \subseteq \Theta_F = \mathbb{R}^k$. Lack of minimality follows form the fact that, since F is a face of $\mathcal{C}_{\mathcal{N}}$, there exists a vector $\boldsymbol{\zeta}_F$ and a constant c_F such that $(\mathbf{z}, \boldsymbol{\zeta}_F) = c_F$, a.e.-$\mu_F$. Then, for any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ in $\mathbb{R}^k$ such that $\boldsymbol{\zeta}_F = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$, $(\mathbf{z}, \boldsymbol{\theta}_1) = (\mathbf{z}, \boldsymbol{\theta}_2) + c_F$, a.e.-$\mu_F$. This, in turn, implies

$$(\mathbf{z}, \boldsymbol{\theta}_1) - \psi_F(\boldsymbol{\theta}_1) = (\mathbf{z}, \boldsymbol{\theta}_2) - \psi_F(\boldsymbol{\theta}_2)$$

and, therefore, $p_{\boldsymbol{\theta}_1} = p_{\boldsymbol{\theta}_2}$, so that $\mathbb{E}_{\boldsymbol{\theta}_1}[\mathrm{A}^*\mathbf{n}] = \mathbb{E}_{\boldsymbol{\theta}_2}[\mathrm{A}^*\mathbf{n}] = \boldsymbol{\xi}$, for some $\boldsymbol{\xi} \in \mathrm{ri}(F)$. Note that, by Hölder's inequality, this is equivalent to the function being no longer strictly convex. We conclude that for any $\boldsymbol{\xi} \in \mathrm{ri}(F)$, $\Lambda_F^{-1}(\boldsymbol{\xi}) := \{\boldsymbol{\theta} \in \Theta_F \colon \nabla\psi_F(\boldsymbol{\theta}) = \boldsymbol{\xi}\}$ is a subset of Θ_F, so that the MLE of $\boldsymbol{\theta}$, if existent, will be an affine subspace in $\mathbb{R}^k$. Nevertheless, despite lack of minimality, each point $\boldsymbol{\xi} \in \mathrm{ri}(F)$ identifies one probability distribution. Explicitly, if $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathrm{ri}(F)$, with $\boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_2$, then $p_{\boldsymbol{\theta}_1} \neq p_{\boldsymbol{\theta}_2}$, for every $\boldsymbol{\theta}_1 \in \Lambda_F^{-1}(\boldsymbol{\xi}_1)$ and every $\boldsymbol{\theta}_2 \in \Lambda_F^{-1}(\boldsymbol{\xi}_2)$.

Let $L(\mathcal{C}_{\mathcal{N}})$ be the face lattice of $\mathcal{C}_{\mathcal{N}}$, i.e. of the set of all faces of $\mathcal{C}_{\mathcal{N}}$, ordered by inclusion. For each $F \in L(\mathcal{C}_{\mathcal{N}})$, define as above the non-minimal exponential family of distributions with convex support F.

Definition 4.4. The union of all such families as F ranges in $L(\mathcal{C}_{\mathcal{N}})$ is called the extended exponential family. (Brown, 1986, denote these families as *aggregate families*.)

Next, since

$$\mathcal{C}_{\mathcal{N}} = \biguplus_{F \in L(\mathcal{C}_{\mathcal{N}})} \mathrm{ri}(F).$$

(see, for example, Ziegler, 1998), where $\biguplus$ denotes union of disjoint sets, for each point $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$ there exists only one face F containing $\boldsymbol{\xi}$ in its relative interior and hence only one sub-family of distributions whose convex support is precisely F. In fact, as noted above, for any face F, the points in $\mathrm{ri}(F)$ define a partition of $\Theta_F = \mathbb{R}^k$ into equivalence classes of affine subspaces $\left\{\Lambda_F^{-1}(\boldsymbol{\xi}), \boldsymbol{\xi} \in \mathrm{ri}(F)\right\}$, each identifying one distribution of exponential type parametrized by $\boldsymbol{\xi}$, i.e. such that $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{z}] = \boldsymbol{\xi}$, for all $\boldsymbol{\theta} \in \Theta_F$.

Combining these considerations, we see that the family of densities of the aggregate exponential family is, under the mean value parametrization,

$$\left\{p_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}, \tag{10}$$

where, for any $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$,

$$p_{\boldsymbol{\xi}}(\mathbf{z}) = \frac{\exp^{(\boldsymbol{\theta}_{\boldsymbol{\xi}}, \mathbf{z})}}{\int \exp^{(\boldsymbol{\theta}_{\boldsymbol{\xi}}, \mathbf{x})} \, d\mu_F(\mathbf{x})},$$

$\boldsymbol{\theta}_{\boldsymbol{\xi}}$ is any point in $\Lambda_F^{-1}(\boldsymbol{\xi})$ and F is the face (possibly improper) of $\mathcal{C}_{\mathcal{N}}$ such that $\mathrm{ri}(F) \ni \boldsymbol{\xi}$. The subset of the densities from (10) given by

$$\left\{p_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})}$$

exhausts the regular linear exponential family of distributions (8) since, for each $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$, the mean value parametrization is a bijection between Θ and $\mathrm{ri}(\mathcal{C}_{\mathcal{N}})$. The extended exponential family is an enlargement of the family (8) obtained by adding distributions, expressible in exponential from, that are parametrized by the points on the boundary of the convex support.

The next Theorem gives regularity properties of this extended family. The proof relies mostly on results due to Brown (1986).

Theorem 4.5. *Assume the extended exponential family with densities as in (10). Then, for any* $\mathbf{z} = \mathrm{A}^*\mathbf{n} \in \mathcal{C}_{\mathcal{N}}$,

- **i.** *The MLE of* $\mathbb{E}[\mathbf{z}]$ *exists always, is unique and is* $\mathbf{z}$*, i.e.* $p_{\mathbf{z}}(\mathbf{z}) = \sup_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}} p_{\boldsymbol{\xi}}(\mathbf{z})$.
- **ii.** $p_{\boldsymbol{\xi}}(\mathbf{z})$ *is a continuous function of* $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$.
- **iii.** $\psi^*(\mathbf{z}) = \sup_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}} p_{\boldsymbol{\xi}}(\mathbf{z})$.

More generally, the theorem holds for any real valued $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$ and not just for the integer-valued sufficient statistics $\mathbf{z}$. The first part of the Theorem shows that the MLE of the mean value is always defined for the extended family and satisfies trivially the moment equations. The second part says that the densities are parametrized continuously by the expectation parameters and the

third statement that the supremum over over the densities (10) coincides with the corresponding value of the sup-log-likelihood function.

Theorem 4.2 and part **iii.** of Theorem 4.5 complement each other. In fact, the former shows that the log-likelihood in a regular exponential family always admits a supremum, attainable in the limit of sequences $\{\boldsymbol{\xi}_i\}$ of points inside the relative interior of the convex support. The latter implies that such a limit is a valid mean vector of the restricted exponential sub-family whose convex support is the face of $\mathcal{C}_{\mathcal{N}}$ containing it in its relative interior.

4.2 Extended Exponential Families for Cell Counts and Extended MLE of the Cell Mean Vector

In this section, we will show that the family (10) corresponds to an extended family of distributions for the cell counts $\mathbf{n}$, where the correspondence is given by a bijection between the mean vectors of the sufficient statistics and the unconditional cell mean vectors. This new family enjoys all the properties of the family (10). In particular, the maximum likelihood estimate of the cell mean vector is always defined.

We first show that the sub-families parametrized by points inside a face F of the convex support corresponds to distribution for cell counts specified by certain improper Poisson sampling (see Definition 2.2).

Lemma 4.6. *Each face F of $\mathcal{C}_{\mathcal{N}}$ corresponds to an improper conditional Poisson model with constraint subspace $\mathcal{N}_F$ such that $\mathcal{N} \subsetneq \mathcal{N}_F \subsetneq \mathcal{M}$ and a set $\mathcal{F} \subsetneq \mathcal{I}$ such that, a.e.-$\nu_{\mathcal{N}_F}$, $\mathbf{n}(i) = 0$ for each $i \in \mathcal{I} \setminus \mathcal{F}$.*

The sets $\mathcal{F}$ corresponding to the faces F, called *facial sets*, were introduced by Geiger et al. (2006). The proof of Lemma 4.6 shows that a subset of $\mathcal{I}$ is a facial set $\mathcal{F}$ for the polyhedral cone $\mathrm{C_A}$ if and only if there exists a vector $\boldsymbol{\zeta}_F \in \mathbb{R}^d$ such that $(\boldsymbol{\zeta}_{\boldsymbol{F}}, \mathbf{a}_i) = 0$ for each $i \in \mathcal{F}$ and $(\boldsymbol{\zeta}_{\boldsymbol{F}}, \mathbf{a}_i) < 0$ for each $i \notin \mathcal{F}$, where $\mathbf{a}_i$ denotes the i-the column of the matrix A. Furthermore, the proof can be reversed to show that, if there exists a set $\mathcal{F} \subsetneq \mathcal{I}$ and a subspace $\mathcal{N} \subsetneq \mathcal{N}_F \subsetneq \mathcal{M}$ such that, a.e.-$\nu_{\mathcal{N}_F}$, $\mathbf{n}(i) = 0$ for each $i \in \mathcal{I} \setminus \mathcal{F}$, then $\mathcal{F}$ identifies a face F of $\mathcal{C}_{\mathcal{N}}$.

By Lemma 3.9, if $\mathcal{F}$ is a facial set corresponding to a face of $\mathcal{C}_{\mathcal{N}}$, then $\mathcal{F}$ is also a facial set for a face of the marginal cone $\mathrm{C_A}$. Then, using facial sets it is possible to obtain a combinatorial restatement of Corollary 3.10 and of Theorem 2.2 in Haberman (1974) (see the end of the Appendix B).

Corollary 4.7. *The MLE does not exist if and only if $\mathrm{supp}(\mathbf{n}) \subseteq \mathcal{F}$ for some facial set $\mathcal{F}$ corresponding to a proper face of the marginal cone $\mathrm{C_A}$.*

The examples in Section 3.1 present facial sets corresponding to facets of the marginal cones for various hierarchical log-linear models. The combinatorial complexity of these marginal cones appear to be rather big, as indicated by the large number of facets associated to tables of small dimensions. See Eriksson et al. (2005) for a combinatorial analysis of the marginal cones corresponding to the hierarchical model of no-3-factor effects under Poisson sampling scheme for various 3-way tables.

We derive a more convenient characterization of the proper faces F of $\mathcal{C}_{\mathcal{N}}$ than the one presented in Lemma 4.6. In fact, each face of $\mathcal{C}_{\mathcal{N}}$ with facial set $\mathcal{F}$ can be described as the convex support for the sufficient statistics of a proper conditional Poisson model defined as in Section 3 by a sub-matrix $\mathrm{A}_{\mathcal{F}}$ of A obtained by considering only the columns of A with indexes in $\mathcal{F}$. The

cells not in $\mathcal{F}$ are then modeled as structural zeros. Formally, let $\pi_{\mathcal{F}} : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{F}}$ be the coordinate projection onto $\mathcal{F}$ given by

$$\pi_{\mathcal{F}}(\mathbf{x}) = \{\mathbf{x}(i), i \in \mathcal{F}\} \tag{11}$$

and notice that $\pi_{\mathcal{F}}(\mathcal{M})$ is the row range of $\mathrm{A}_{\mathcal{F}}$. This argument is made rigorous in the next result.

Corollary 4.8. *Each face F of $\mathcal{C}_{\mathcal{N}}$ is the convex support of a proper conditional Poisson model with design matrix $\mathrm{A}_{\mathcal{F}}$ and constraint subspace $\pi_{\mathcal{F}}(\mathcal{N})$.*

Therefore, the distributions in (10) corresponding to faces of the convex support can equivalently be described using proper conditional Poisson schemes with structural zero or improper schemes in which additional sampling constraints force some of the cells counts to be zero. The former representation will be adopted here for convenience, although the latter provide a better interpretation for the extended MLE of the cell mean vector as a non-negative vector with data-dependent support.

We now proceed with the construction of the extended family for the cell counts by defining densities for the cell counts parametrized by $\mathcal{C}_{\mathcal{N}}$. We show that there exists a one-to-one correspondence between the points in $\mathcal{C}_{\mathcal{N}}$ and the unconditional cell mean vectors arising from the log-linear models $\pi_{\mathcal{F}}(\mathcal{M})$, with $F \in L(\mathcal{C}_{\mathcal{N}})$. To this end, we assume initially Poisson sampling, so that convex support is the marginal cone C_{A}, where A is full row-rank (this assumption is inessential, but it simplifies the arguments below). We distinguish two cases:

1. If $\boldsymbol{\xi} \in \mathrm{ri}(\mathrm{C}_{\mathrm{A}})$, the vector $\boldsymbol{\theta}_{\boldsymbol{\xi}} = \Lambda_F^{-1}(\boldsymbol{\xi}) \in \mathbb{R}^k$ is unique, with $k = \dim(\mathcal{M})$. Set $\boldsymbol{\mu}_{\boldsymbol{\xi}} = (\mathrm{A})^{\top}\boldsymbol{\theta}_{\boldsymbol{\xi}}$ and $\mathbf{m}_{\boldsymbol{\xi}} = \exp\{\boldsymbol{\mu}_{\boldsymbol{\xi}}\}$, so that $\mathrm{supp}(\mathbf{m}_{\boldsymbol{\xi}}) = \mathcal{I}$. Since $\boldsymbol{\mu}_{\boldsymbol{\xi}} \in \mathcal{M}$, Proposition 2.1 implies that $\mathbf{m}_{\boldsymbol{\xi}}$ is a mean vector for the contingency table $\mathbf{n}$ corresponding to the model (15). Moreover, $\mathbf{m}_{\boldsymbol{\xi}}$ satisfies the moment equations $\boldsymbol{\xi} = \mathrm{A}\mathbf{m}_{\boldsymbol{\xi}}$. Conversely, if $\mathbf{m}$ is a mean vector for the log-linear model $\mathcal{M}$, there exists a unique $\boldsymbol{\xi} \in \mathrm{ri}(\mathrm{C}_{\mathrm{A}})$ such that $\boldsymbol{\xi} = \mathrm{A}\mathbf{m}$.

2. The argument carries over almost unchanged if $\boldsymbol{\xi} \in \mathrm{ri}(F)$, for some face F with facial set $\mathcal{F}$. Consider the log-linear subspace $\pi_{\mathcal{F}}(\mathcal{M})$ spanned by the rows of the sub-matrix $\mathrm{A}_{\mathcal{F}}$. Then everything takes place within this reduced exponential family for contingency tables defined over the cells in $\mathcal{F}$. Specifically, each $\boldsymbol{\xi} \in \mathrm{ri}(F)$ corresponds to a unique strictly positive cell mean vector $\mathbf{m}_{\boldsymbol{\xi}} \in \mathbb{R}^{\mathcal{F}}$ such that $\boldsymbol{\xi} = \mathrm{A}_{\mathcal{F}}\mathbf{m}_{\boldsymbol{\xi}}$ and vice versa, for every mean vector $\mathbf{m}$ whose logarithm belongs to $\pi_{\mathcal{F}}(\mathcal{M})$, there exists a unique point $\boldsymbol{\xi} \in \mathrm{ri}(F)$ such that $\boldsymbol{\xi} = \mathrm{A}_{\mathcal{F}}\mathbf{m}$, where $\mathrm{A}_{\mathcal{F}}$ is a (now non full rank) matrix whose rows span $\pi_{\mathcal{F}}(\mathcal{M})$, with $\mathrm{rank}(\mathrm{A}_{\mathcal{F}}) = \dim(\pi_{\mathcal{F}}(\mathcal{M}))$.

We proved in Lemma 3.9 that the points of the corresponding convex support $\mathcal{C}_{\mathcal{N}}$ are mapped injectively by the linear map f_{A} into C_{A} and that the facial sets for $\mathcal{C}_{\mathcal{N}}$ are also facial sets for C_{A}. Then, for each $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$, we will write $\mathbf{m}_{\boldsymbol{\xi}}$ for the nonnegative vector determined by $f_{\mathrm{A}}(\boldsymbol{\xi}) \in \mathrm{C}_{\mathrm{A}}$, in the sense that $\mathrm{supp}(\mathbf{m}_{\boldsymbol{\xi}}) = \mathcal{F}$ is the appropriate facial set for $\mathcal{C}_{\mathcal{N}}$ and $f_{\mathrm{A}}(\boldsymbol{\xi}) = \mathrm{A}_{\mathcal{F}}\mathbf{m}$. Existence and uniqueness of $\mathbf{m}_{\boldsymbol{\xi}}$ follow from the above considerations regarding the Poisson case. Therefore, for the sampling subspace $\mathcal{N}$, the extended exponential family of distributions for the cell counts $\mathbf{n}$ with densities

$$\left\{ p_{\mathbf{m}_{\boldsymbol{\xi}}} \right\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}, \tag{12}$$

is well defined, where

$$p_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{x}) = \frac{\exp^{(\boldsymbol{\theta}_{\boldsymbol{\xi}}, \mathrm{A}_{\mathcal{F}}^{*}\mathbf{x})}}{\int_{\mathbb{R}^{\mathcal{I}}} \exp^{(\boldsymbol{\theta}_{\boldsymbol{\xi}}, \mathrm{A}_{\mathcal{F}}^{*}\mathbf{x})} \, d\nu_{\mathcal{N}_F}(\mathbf{x})},$$

with $\boldsymbol{\xi} \in \mathrm{ri}(F)$, $\boldsymbol{\theta}_{\boldsymbol{\xi}} \in \Lambda_F^{-1}(\boldsymbol{\xi})$ and $\mathcal{F}$ is the facial set corresponding to F .

Note that the distributions parametrized by points $\mathbf{m}_{\boldsymbol{\xi}}$, with $\boldsymbol{\xi} \in \mathrm{ri}(F)$, are, by construction, defined only over $\mathbb{R}^{\mathcal{F}}$. It is more convenient to have them defined over the whole $\mathbb{R}^{\mathcal{I}}$ instead. For this purpose, consider, for every facial set $\mathcal{F}$,the log-partition function $\tau_{\mathcal{F}} : \mathbb{R}^{\mathcal{F}} \to \mathbb{R}^{\mathcal{I}}$ given by

$$\tau_{\mathcal{F}}(\mathbf{x}) = \begin{cases} \mathbf{x}(i) & i \in \mathcal{F} \\ 0 & i \in \mathcal{I} \backslash \mathcal{F}. \end{cases} \tag{13}$$

For $\boldsymbol{\xi} \in \mathrm{ri}(F)$, we will identify $\mathbf{m}_{\boldsymbol{\xi}}$ with $\tau_{\mathcal{F}}(\mathbf{m}_{\boldsymbol{\xi}})$, so that the corresponding distribution in (12) is defined over $\mathbb{R}^{\mathcal{I}}$, with the understanding that the cells not in $\mathcal{F}$ are to be treated like structural zeros, hence not affecting the likelihood.

The regularity properties of the extended family of distributions for the sufficient statistics from the previous section derive essentially from the topological properties of the polyhedron $\mathcal{C}_{\mathcal{N}}$. The next theorem shows that such properties are in fact preserved for the set of vectors $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}$ parametrizing the family (12). For an explicit geometric representation of the set $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}$, and for a different proof of the next result, see Section 5.

Theorem 4.9. *The sets $\mathcal{C}_{\mathcal{N}}$ and $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}$ are homeomorphic.*

As a result, the families (12) and (10) have identical properties.

Theorem 4.10. *Assume the extended exponential family with densities as in (12). Then for any $\mathbf{n}$,*

- **i.** *The MLE $\widehat{\mathbf{m}}$ of the conditional cell mean always exists, is unique and satisfies the moment equations $\mathrm{A}^* \widehat{\mathbf{m}} = \mathbf{z} = \mathrm{A}^* \mathbf{m}_{\mathbf{z}}$. Furthermore, $p_{\mathbf{m}_{\mathbf{z}}}(\mathbf{n}) = \sup_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}} p_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{n})$.*
- **ii.** *$p_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{n})$ is a continuous function of $\mathbf{m}_{\boldsymbol{\xi}}$, $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$.*
- **iii.** *$\psi^*(\mathbf{z}) = \sup_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}} p_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{n})$.*

As it was the case with Theorem 4.5, the above results hold for any $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$. In summary, using Proposition 2.1 and Theorem 4.10, it is easy to see that the set of vectors $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}$ satisfies two defining conditions:

$$\begin{aligned} &\boldsymbol{\xi} = \mathrm{A}^* \mathbf{m}_{\boldsymbol{\xi}}, \\ &\ln \pi_{\mathcal{F}}(\mathbf{m}_{\boldsymbol{\xi}}) \in \pi_{\mathcal{F}}(\mathcal{M}), \end{aligned} \tag{14}$$

where $\mathcal{F}$ is the facial set for the face F of $\mathcal{C}_{\mathcal{N}}$ for which $\mathrm{ri}(F) \ni \boldsymbol{\xi}$.

The construction carried out so far leads naturally to the following definition of Extended MLE.

Definition 4.11. For any observed value of the sufficient statistics $\mathbf{z}$, $\widehat{\mathbf{m}}^{\mathrm{e}} = \mathbf{m}_{\mathbf{z}}$ is the *Extended Maximum Likelihood Estimate* of $\mathbf{m}_{\boldsymbol{\xi}}$.

Remarks.

a) The parametrization used to construct the family (12) is based on the unconditional cell mean vectors $\mathbf{m}_{\boldsymbol{\xi}}$, i.e. the cell mean vectors arising from the Poisson sampling. As pointed out in the remark following Theorem 3.7, for general conditional Poisson schemes, the conditional cell mean vectors, and consequently their MLEs, may not belong to the set $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}}$. In fact,

the MLEs of the conditional cell mean vectors satisfy the moment equation, as shown in part **i.** of Theorem 4.10, but not necessarily the second condition in Equation (14). Nevertheless, Theorems 4.9 and 4.10 combined show that, in a conditional Poisson model, the MLE of $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu}$, if existent, identifies one $\boldsymbol{\mu} \in \mathcal{M}$ such that $\exp \boldsymbol{\mu} = \mathbf{m}_{\mathbf{z}}$, with $\mathbf{z}$ being the observed sufficient statistic. When $\mathbf{z}$ belongs to the boundary of the convex support, then, with respect to the extended family, the MLE of $\mathrm{P}_{\pi_{\mathcal{F}}(\mathcal{M}\ominus\mathcal{N})}\pi_{\mathcal{F}}(\boldsymbol{\mu})$ always exists, is unique and identifies one point $\boldsymbol{\mu} \in \pi_{\mathcal{F}}(\mathcal{M})$ such that $\exp^{\tau_{\mathcal{F}}(\boldsymbol{\mu})} = \mathbf{m}_{\mathbf{z}}$, where $\mathcal{F}$ is the facial set determined by $\mathbf{z}$. Equivalently, the matrix H of Lemma 3.8 indirectly defines a bijection between $\mathcal{M}$ and $\mathcal{M} \ominus \mathcal{N}$, because the sampling constraints fix the parameters determining $\mathcal{N}$. Furthermore, the extended MLE $\mathbf{m}_{\mathbf{z}}$ is *not* in general the MLE of the unconditional mean of the extended family, since the sampling constraints do not allow to estimate the entire parameter space, but only the portion associated to $\mathcal{M} \ominus \mathcal{N}$. By the same token, within the extended family, if $\mathrm{P}_{\mathcal{N}}\mathbf{m}$ can be determined, then the MLE of $\mathbf{m}$ could be recovered from the MLE of $\mathrm{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu}$ (see also Haberman, 1974, Theorem 2.6).

b) For the special cases of product-multinomial and multinomial-Poisson (Lang, 2004) sampling schemes, the extended MLE is the MLE of the conditional mean vector, with respect to the extended family, and it also coincides with the extended MLE of the unconditional mean. In fact, it is easy to see that Theorem 3.7 holds true for the extended MLE as well. For these sampling schemes, the definition of extended MLE proposed here generalizes the notion of extended MLE for the cell mean vector originally suggested by Haberman (1974) and further developed by Fienberg et al. (1980) and Lauritzen (1996). In particular, the set $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi}\in\mathcal{C}_{\mathcal{N}}}$ can be more conveniently thought as the limit closure of all possible positive cell mean vectors. Then, for any observed table $\mathbf{n}$ and any sequence $\{\boldsymbol{\xi}_n\} \subset \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ such that $\mathrm{A}^*\mathbf{n} = \lim_n \boldsymbol{\xi}_n$, one can equivalently compute the extended MLE $\widehat{\mathbf{m}}^{\mathrm{e}}$ as the limit

$$\widehat{\mathbf{m}}^{\mathrm{e}} = \lim_n \mathbf{m}_{\boldsymbol{\xi}_n},$$

where each $\mathbf{m}_{\boldsymbol{\xi}_n}$ is a positive cell mean vector. This is precisely how the iterative proportional fitting algorithm works (see discussion at the end of the next Section 5).

c) By construction, the extended MLE satisfies the moment equation, namely $\mathrm{P}_{\mathcal{M}}\widehat{\mathbf{m}}^{\mathrm{e}} = \mathrm{P}_{\mathcal{M}}\mathbf{n}$, a feature proved by Fienberg et al. (1980) and Lauritzen (1996), in less generality.

We conclude this section by showing that the extended exponential family for cell counts arises in a natural way as the closure of the regular exponential family with respect to the total variation and the reverse information metric. See Csiszár and Matúš (2003, 2005) for an exhaustive account and generalization of these notions of closure. Determining the point $\mathbf{m}_{\mathbf{z}}$ corresponds to computing a "reverse" Kullback-Lieber projection onto the set of all probability measures which are absolutely continuous with respect to the base measure $\mu_{\mathcal{N}}$ and whose mean parameter satisfies the moment equations. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures defined on the same probability space. The I-divergence of P from Q is defined as

$$D(\mathbb{P}||\mathbb{Q}) = \begin{cases} \int \ln \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{P} & \text{if} \quad \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{otherwise,} \end{cases}$$

and the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ is a metric $||\cdot||_{\mathrm{tv}}$ given by

$$||\mathbb{P} - \mathbb{Q}||_{\mathrm{tv}} = 2 \sup_B |\mathbb{P}(B) - \mathbb{Q}(B)|,$$

(a)

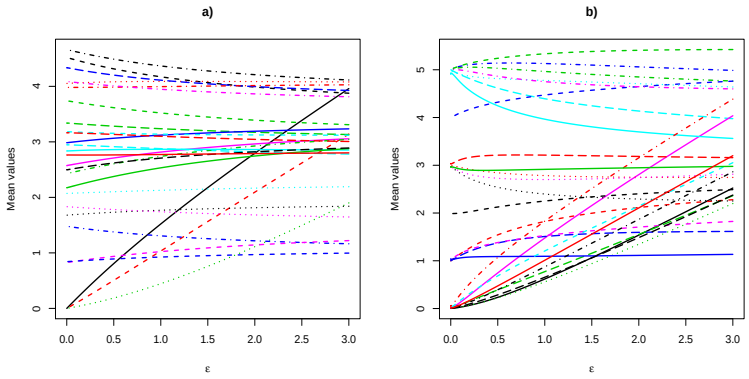
ϵ	1	5	4	2	3	5	1	2
ϵ	2	5	3	5	2	3	2	3
ϵ	4	1	1	2	3	2	4	4

(b)

ϵ	5	2	1	3	1	ϵ	4	ϵ
1	5	3	5	ϵ	ϵ	3	5	ϵ
ϵ	ϵ	3	1	ϵ	3	3	5	5

Table 1: 3^3 table and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$. **(a):** The ϵ-cells correspond to a null margin. **(b):** The ϵ-cells determine a co-facet for the corresponding marginal cone (see example **b)** in Section 3.1).

Figure 1: Sequence $\{\mathbf{m}_{\boldsymbol{\xi}_\epsilon}\}$ of the cell mean vectors as $\epsilon \to 0$ for the Tables 1**(a)** and 1**(b)**, in part **a)** and **b)**, respectively.



where the supremum is taken over all measurable Borel sets B. A sequence $\{\mathbb{Q}_n\}$ of probability measures is said to rI-converge to a probability measure $\mathbb{P}$ if

$$\lim_n D(\mathbb{P}||\mathbb{Q}_n) = 0,$$

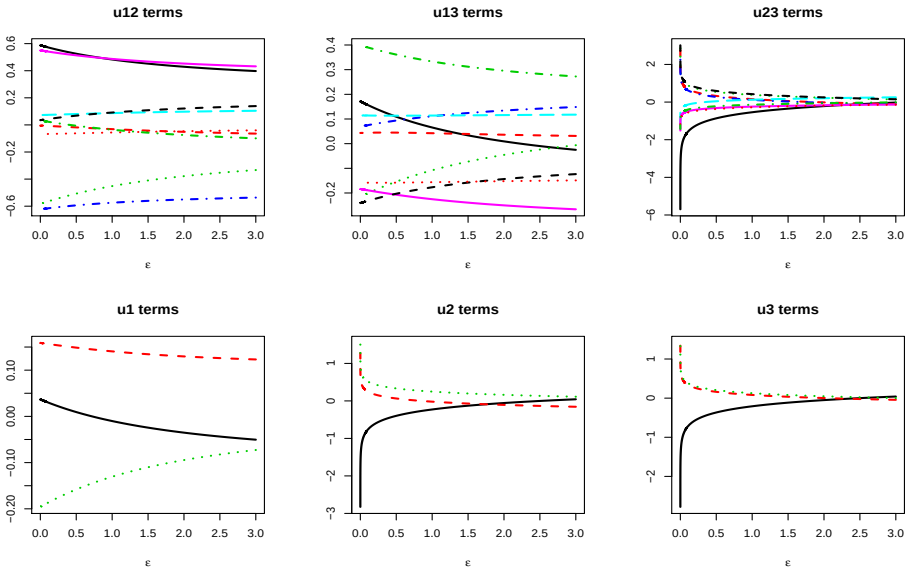
while is said to converge in total variation when

$$\lim_n ||\mathbb{Q}_n - \mathbb{P}||_{\text{tv}}.$$

Theorem 4.12. *The family (12) is closed with respect to the rI diverge and total variation metric. In fact, it is the rI-closure and the and total variation closure of the family* $\left\{p_{\mathbf{m}_{\boldsymbol{\xi}}}\right\}_{\boldsymbol{\xi}\in \text{ri}(\mathcal{C}_{\mathcal{N}})}$.

Example 4.13 (The Extended MLE). This example exemplifies the nature of the extended MLE as a point in the sequential closure of the set $\{\exp^{\boldsymbol{\mu}}\colon \boldsymbol{\mu}\in\mathcal{M}\}$. Assume the model of no-3-factor effect

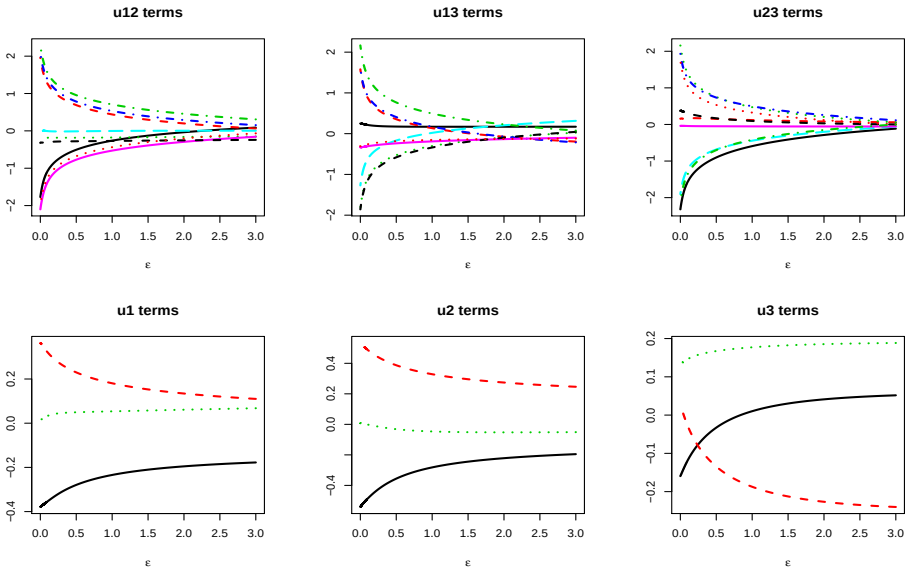
Figure 2: Sequences of u-term expansions of the points in $\mathbf{m}_\epsilon$ as $\epsilon \to 0$ for the Table 1**(a)** and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$ of Example 4.13.



$\{\{1,2\},\{2,3\},\{1,3\}\}$ and Poisson sampling for the 3^3 contingency tables of Tables 1**(a)** and 1**(b)**. Both tables contain positive integer entries (randomly generated from a uniform distribution on $\{1,2,\ldots,5\}$) except along patterns of cells corresponding to two different co-facets of the marginal cone, which contain instead the same positive real number ϵ. (A co-facet is the complement in $\mathcal{I}$ of facial set corresponding to a facet.) The number ϵ is then let decrease monotonically to zero. In both cases, for every positive ϵ, the corresponding table margins define a point $\boldsymbol{\xi}_\epsilon$ inside the relative interior of $\mathrm{C_A}$. As $\epsilon \downarrow 0$, the sequence $\{\boldsymbol{\xi}_\epsilon\}$ tends to the point on the appropriate facet representing the integer-valued margins for which the MLE is not defined.

For every $\epsilon > 0$, the corresponding vector $\boldsymbol{\xi}_\epsilon$ identifies, in a mean value sense, one probability distribution for the cell counts with cell mean vector $\mathbf{m}_{\boldsymbol{\xi}_\epsilon}$. In the limit, the MLE does not exist, but the extended MLE, which is $\lim_{\epsilon\downarrow 0} \mathbf{m}_{\boldsymbol{\xi}_\epsilon}$, is well defined. Figure 1 shows the two sequences $\{\mathbf{m}_{\boldsymbol{\xi}_\epsilon}\}_\epsilon$ for the Tables 1**(a)** and 1**(b)**, as a function of ϵ. In both cases, as the margins approach the boundary of the marginal cone, the values of the coordinates of $\mathbf{m}_\epsilon$ defining the co-facet tend to 0 in a continuous fashion. In contrast, Figures 2 and 3 show the sequences of the u-terms (see, for example, Bishop et al., 1975) for the expansions of the points $\{\log \mathbf{m}_\epsilon\}_\epsilon$, for the Tables 1**(a)** and 1**(b)**, respectively. It is easy to see that some of the u-terms explode to infinity as ϵ approaches 0, an indication of the fact that they cannot be estimated. The rate at which the diverging terms tend to infinity is $\frac{1}{\epsilon}$. It is interesting to point out that for Table 1**(a)**, among the u-terms of highest order, only some of the $u_{(2,3)}$-terms diverge. Because of the hierarchical nature of the model, this discontinuity at 0 also affects the lower order terms $u_{(2)}$ and $u_{(3)}$ as well. It is immediate to see why only the $u_{(2,3)}$ terms are involved: the zero margin, achieved in the limit, is one of the $\{2,3\}$ marginal configurations. The computations were performed in `R` (R Development Core Team, 2005) using the `loglin` routine.

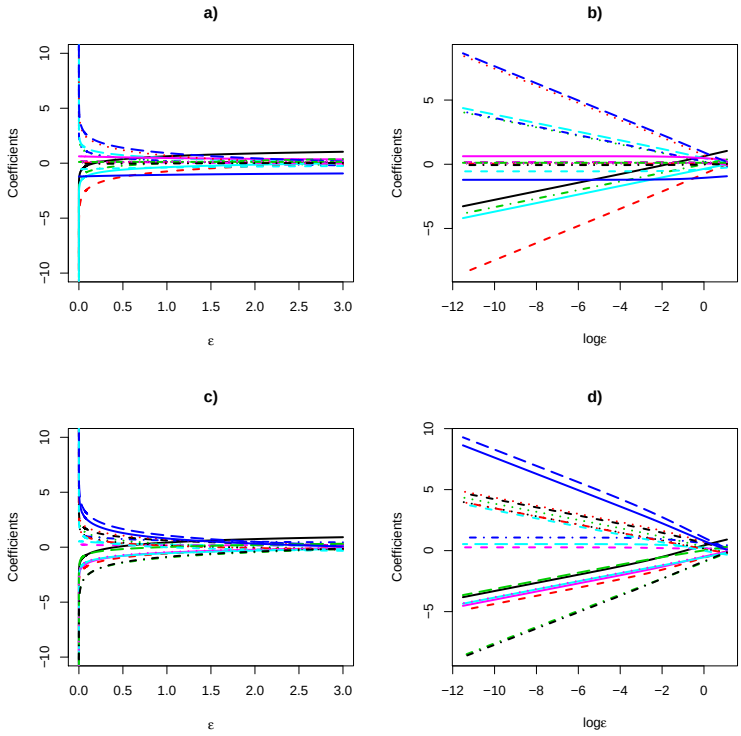
Figure 3: Sequences of u-term expansions of the points in $\mathbf{m}_\epsilon$ as $\epsilon \to 0$ for the Table 1**(b)** and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$ of Example 4.13.



The explosive behavior of the estimates of u-terms when the MLE is non-existent takes place also with the estimates of natural parameters. In particular, Lemma 7.1 in the Appendix A illustrates why the Newton-Raphson procedure for maximizing the log-likelihood function, utilized for fitting generalized linear models, can be affected by numerical instabilities. In fact, as the log-likelihood approaches its supremum, some of the coordinates of the estimated vector of natural parameters will necessarily explode. These directions of recession of the log-likelihood function are determined by the normal cone at the observed sufficient statistics to the supporting hyperplane for the face of the marginal cone containing it (see Lemma 7.2 in the Appendix A). Figure 4 displays the sequences of natural parameters for the Tables 1**(a)** and 1**(b)** versus both ϵ and $\log \epsilon$ (we chose to plot the estimates also on the log scale to improve the readability). Like with the u-terms parametrization, some parameters in the linear expansion of the logarithm of the sequence $\{\mathbf{m}_\epsilon\}_\epsilon$ diverge to infinity as $\epsilon \downarrow 0$. In fact, inspection of the hessian of both the Poisson and product-multinomial log-likelihood functions reveals that, as the algorithm progresses, they become closer to be singular (see Fienberg and Rinaldo, 2006a), thus causing potential numerical instabilities. In this example, the natural parameters were computed using the `glm` routine in `R` with the parameter `family` set to `poisson`. (The weighted least square procedure implemented in the `glm` routine is a specialized version of the Newton-Raphson procedure). The full rank design matrix was also computed in `R`, using sum-zero contrasts.

Note that the previous examples can be carried out in more elaborated settings in which the ϵ-cells are replaced by different sequences vanishing, not necessarily in a monotone fashion, at different rates. The conclusions would remain unchanged. ■

Figure 4: Sequence of the natural parameter coefficients computed by the `glm` routine for the points in $\mathbf{m}_\epsilon$ versus ϵ and $\log \epsilon$ for the Table 1**(a)** (parts **a)** and **b)**, respectively) and Table 1**(b)** (part **c)** and **d)**, respectively) for and the model $\{\{1,2\},\{2,3\},\{1,3\}\}$.



5 Extended Exponential Families and Toric Varieties

In this section we will derive a geometric characterization of the mean vectors $\left\{\mathbf{m}_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi}\in \mathrm{C_A}}$, where $\mathrm{C_A}$ is the marginal cone determined by an integer-valued design matrix A not necessarily of full rank. It will be convenient to think of $\mathrm{C_A}$ as the convex support associated with the sufficient statistics for the log-linear model specified by A under Poisson sampling scheme. The assumption that A has integer entries is hardly restrictive, as the most common log-linear models are in fact defined in this way. The matrix A determines a monomial map $\phi_{\mathrm{A}} : \mathbb{R}^d_{>0} \mapsto \mathbb{R}^{\mathcal{I}}_{>0}$ given by

$$\mathbf{z} = (z_1, \ldots, z_d) \mapsto \left(\prod_{j=1}^{d} z_j^{a_{j,1}}, \ldots, \prod_{j=1}^{d} z_j^{a_{j,|\mathcal{I}|}} \right) = (\mathbf{z}^{\mathbf{a}_1}, \ldots, \mathbf{z}^{\mathbf{a}_{|\mathcal{I}|}}), \tag{15}$$

where $a_{j,l}$ denotes the (j,l)-th element of A. The terminology "monomial map" is appropriate since the image of ϕ_{A} is a positive vector whose coordinates are obtained by evaluating monomial expressions. The relationship between the linear exponential family (2) and the monomial map (15) is straightforward and is given in in the following lemma.

Lemma 5.1. *The following facts hold true:*

i. $\log\left(\mathrm{im}(\phi_{\mathrm{A}})\right) = \mathcal{M}$.

ii. The family of distributions $\{P_{\boldsymbol{\eta}}\}_{\boldsymbol{\eta}\in\mathbf{E}}$ *is the linear exponential family as in (2) if and only if, for every* $\boldsymbol{\eta} \in \mathbf{E}$, $p_{\boldsymbol{\eta}} c_{\boldsymbol{\eta}} \in \mathrm{im}(\phi_{\mathrm{A}})$, *for some constant* $c_{\boldsymbol{\eta}}$ *depending on* $\boldsymbol{\eta}$.

As a result, for any $\boldsymbol{\mu} \in \mathcal{M}$, $\mathbf{m} = \exp^{\boldsymbol{\mu}} \in \mathrm{im}(\phi_{\mathrm{A}})$. That is, the image of the monomial map (15) is the set of all positive cell mean vectors. Statistical models described by monomial equations as in (15) are called *toric models*, a terminology introduced by Geiger et al. (2006).

The following example shows how monomial maps offer a different, equivalent representation of hierarchical log-linear models (see, for example, Darroch and Speed, 1983). The generalization to any hierarchical models is immediate.

Example 5.2. Consider the 2^3 table and the decomposable model $\Delta = \{\{1,2\},\{2,3\}\}$. The cell set $\mathcal{I} = \{1,2\} \otimes \{1,2\} \otimes \{1,2\}$ is linearized according to Equation (1), and the 8×8 matrix

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

used in the map (15), gives

$$\phi_{\mathrm{A}}(\mathbf{z}) = (\mathbf{z}^{\mathbf{a}_1}, \ldots, \mathbf{z}^{\mathbf{a}_{|\mathcal{I}|}}) = (z_1 z_5, z_1 z_6, z_2 z_7, z_2 z_8, z_3 z_5, z_3 z_6, z_4 z_7, z_4 z_8)\,, \tag{16}$$

where the coordinates of $\mathbf{z}$ are ordered according to the row ordering of A, shown in Table 2 and determined using Equation (1).

d	i_d	z_i
$\{1,2\}$	11	z_1
$\{1,2\}$	12	z_2
$\{1,2\}$	21	z_3
$\{1,2\}$	22	z_4
$\{2,3\}$	11	z_5
$\{2,3\}$	12	z_6
$\{2,3\}$	21	z_7
$\{2,3\}$	22	z_8

Table 2: Row ordering for the matrix A in Example 5.2.

Let F_d denote the subspace of $\mathbb{R}^{\mathcal{I}}$ consisting of the set of functions that depends on $i \in \mathcal{I}$ only through i_d, $d \in \Delta$ (i.e. $f(i) = f(j)$ if and only if $i_d = j_d$). For any $\mathbf{p} \in \mathrm{im}(\phi_{\mathrm{A}})$, from (16) it follows that

$$\log(\mathbf{p}) = \sum_{d\in\Delta} f_d(i_d), \tag{17}$$

where each function f_d belongs to the corresponding subspace F_d, $d \in \Delta$. Equation 17 implies that $\log(\mathbf{p})$ belongs to the factor-interaction subspace defined by the decomposable generating class Δ, in the terminology of Darroch and Speed (1983) and Lauritzen (1996). ■

Monomial maps allow for a parametrization of the set of positive probability distributions defined by the log-linear subspace $\mathcal{M}$ that is alternative to the one specified by the corresponding linear exponential family. The key difference between these two approaches is that, while linear exponential families are structurally linked to log-linear subspaces, monomial maps produce instead direct representations of the distributions of interest. More importantly, the set $\mathrm{im}(\phi_{\mathrm{A}})$ can be characterized geometrically using the solution set of a system of polynomials equations (see, in particular, Cox et al., 1996; Sturmfels, 1996; Diaconis and Sturmfels, 1998). For each $i \in \mathcal{I}$ introduce the indeterminate $\mathrm{x_i}$ in the ring of polynomial equations $k[\mathbf{x}]$, where the field k in the present context can be taken to be $\mathbb{R}$. Consider the lattice $\mathcal{L}_{\mathrm{A}} = \mathrm{kernel}(\mathrm{A}) \cap \mathbb{Z}^{\mathcal{I}}$ and the system of polynomial equations

$$I_{\mathrm{A}} := \langle \mathbf{x}^{\mathbf{u}^+} - \mathbf{x}^{\mathbf{u}^-} : \mathbf{u} \in \mathcal{L}_{\mathrm{A}} \rangle,$$

where $\mathbf{u}^+ = \{\max(\mathbf{u}(i), 0), i \in \mathcal{I}\}$ and $\mathbf{u}^- = \{-\min(\mathbf{u}(i), 0), i \in \mathcal{I}\}$ The ideal I_{A} is called *toric ideal* (Sturmfels, 1996). The set of non-negative real solution of the polynomial system I_{A} is the irreducible affine *toric variety*

$$V_{\mathrm{A},\geq 0} = \mathrm{V}(\mathcal{I}_{\mathrm{A}}) \cap \mathbb{R}^{\mathcal{I}}_{\geq 0}.$$

The defining condition satisfied by all non-negative real valued points of $V_{\mathrm{A},\geq 0}$ is

$$\mathbf{m} \in V_{\mathrm{A},\geq 0} \Longleftrightarrow \mathbf{m}^{\mathbf{u}^+} = \mathbf{m}^{\mathbf{u}^-}, \quad \forall \mathbf{u} \in \mathcal{L}_{\mathrm{A}} \tag{18}$$

The following examples, along with others which can be found, for example in Diaconis and Sturmfels (1998), Geiger et al. (2006) and Pachter and Sturmfels (2005), illustrate this polynomial representation.

Example 5.3. Consider the 2×2 table with cell mean vector, in tabular notation,

m_{11}	m_{12}
m_{21}	m_{22}

and the model of independence $\Delta = \{\{1\}, \{2\}\}$, with the usual cell ordering as in Example 3.6. The 3-dimensional log-linear subspace $\mathcal{M}$ is spanned by the 4 vectors defining the row and column sums (see Example 3.6 above), which are

1	1
0	0

, | | |
|---|---|
| 0 | 0 |
| 1 | 1 |

, | | |
|---|---|
| 1 | 0 |
| 1 | 0 |

and | | |
|---|---|
| 0 | 1 |
| 0 | 1 |

.

These 4 spanning vectors are the rows of the associated design matrix A in Equation (7). The 1-dimensional orthogonal subspace of $\mathcal{M}$ in $\mathbb{R}^4$ is spanned by the vector

$+1$	-1
-1	$+1$

.

The toric ideal for this model is the principal ideal $I_{\mathrm{A}} = \langle \mathrm{x}_1\mathrm{x}_4 - \mathrm{x}_2\mathrm{x}_3 \rangle$. It is immediate to see that the associated toric variety

$$V_{\mathrm{A},\geq 0} = \left\{ \mathbf{m} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \colon m_{11}m_{22} = m_{12}m_{21} \right\}. \tag{19}$$

The variety $V_{\mathrm{A},\geq 0}$ is known in algebraic geometry as Segre variety and its restriction to the interior of the simplex in $\mathbb{R}^4$ is renown in statistics as the surface of independence(Fienberg and Gilbert, 1970). In fact, points in the (interior of the) simplex satisfying Equation (19) are the set of all distributions on the set $\mathcal{I}$ with odds ratio equal to 1, i.e.

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = 1. \quad \blacksquare \tag{20}$$

Example 5.4. For a more sophisticated example, consider a 2^3 table and the model $\Delta = \{\{1,2\},\{2,3\},\{1,3\}\}$ with cell mean vector

m_{111}	m_{121}
m_{211}	m_{221}

m_{112}	m_{122}
m_{212}	m_{222}

.

The dimension of the log-linear subspace $\mathcal{M}$ is 7 and its orthogonal complement in $\mathbb{R}^{\mathcal{I}}$ is spanned by the one vector

$+1$	-1
-1	$+1$

-1	$+1$
$+1$	-1

.

Using this vector and Equation (18), it can be seen that the corresponding toric variety for this model is

$$V_{\mathrm{A},\geq 0} = \{\mathbf{m} \in \mathbb{R}^{\mathcal{I}} \colon m_{111}m_{221}m_{122}m_{212} = m_{121}m_{211}m_{112}m_{222}\},$$

The positive points in this variety will satisfy Equation (3.3-12) of Bishop et al. (1975) for the hypothesis of no-three-factor effect follows,

$$\frac{m_{ijk}m_{rsk}}{m_{rjk}m_{isk}} = \frac{m_{ijt}m_{rst}}{m_{rjt}m_{ist}} \tag{21}$$

for $i \neq r$, $j \neq s$, $k \neq t$. $\blacksquare$

The previous examples show that every positive cell mean vector for the log-linear subspace $\mathcal{M}$ lies in $V_{\mathrm{A},\geq 0}$, an immediate consequence of Equation(18). The traditional log-linear settings hinge upon representing $\log \mathbf{m}$ as a point in the vector space spanned by the rows of A. Despite its apparent simplicity, this approach is severely limited by the constraint that all permissible points $\mathbf{m}$ must be strictly positive. In the above examples, only strictly positive cell mean vector can, in fact, satisfies Equations (20) and (21). In contrast, distributions for which, for example, $m_{11} = m_{21} = 0$, will satisfy (19). In fact, the toric variety representation enjoys the crucial advantage of naturally providing an explicit representation of the closure of the parameter space. This closure consists of of points that belong to the toric variety and have some zero coordinates. It is the possibility of identifying these points, both analytically (polynomials are continuous function) and geometrically, that allows for a full description of all possible patterns of sampling zeros leading to a nonexistent MLE and the definition of extended exponential family and extended MLE. Geiger et al. (2006) proved this far-reaching results that hypersurface $V_{\mathrm{A},\geq 0}$ is the sequential closure of the open set $\mathrm{im}(\phi_{\mathrm{A}})$.

Theorem 5.5 (Geiger et al. (2006)). $\mathrm{cl}(\mathrm{im}(\phi_{\mathrm{A}})) = V_{\mathrm{A},\geq 0}$.

The next logical step is to verify that $V_{\mathrm{A},\geq 0}$, being the closure of $\mathrm{im}(\phi_{\mathrm{A}})$, indeed parametrizes the extended family (12). This is shown in Theorem 5.8, whose formulation and proof rely on some other geometric quantities that are now introduced.

The equivalence in Equation (18) is used to show the following combinatorial result, which demonstrates that, for each $\mathbf{m} \in V_{\mathrm{A},\geq 0}$, $\mathrm{supp}(\mathbf{m})$ is a facial set of A and, conversely, each facial set of A is the support set of points in $V_{\mathrm{A},\geq 0}$.

Lemma 5.6. *For any $\mathbf{m} \in V_{\mathrm{A},\geq 0}$, $\mathrm{supp}(\mathbf{m})$ is a facial set of A and, conversely, for any facial set $\mathcal{F}$ of A, there exist points in $V_{\mathrm{A},\geq 0}$ with $\mathrm{supp}(\mathbf{m}) = \mathcal{F}$.*

For any $\boldsymbol{\xi} \in \mathrm{C}_{\mathrm{A}}$, consider the polyhedron

$$\mathrm{P}_{\boldsymbol{\xi}} = \{\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} : \mathrm{A}\mathbf{x} = \boldsymbol{\xi}\}. \tag{22}$$

For any integer $\mathbf{t} \in \mathrm{C}_{\mathrm{A}}$, the integer points in $\mathrm{P}_{\mathbf{t}}$, called the *fiber* of $\mathbf{t}$, are the set of all tables with the same margins $\mathbf{t}$, i.e. the set of tables in the support of the conditional distribution of $\mathbf{n}$ given the observed statistics. The polyhedron (22) is, in most situation, bounded, i.e. it is a polytope, but this does not necessarily hold for general log-linear models.

Example 5.7. Consider a 3^3 contingency tables for the variables $1, 2$ and 3 under Poisson sampling and the non-hierarchical log-linear model specifying just the interaction $\{1, 2\}$ and the main effect $\{3\}$. The transpose of the design matrix A^* for this model can be chosen to be of the form $[\mathrm{A}^*_1|\mathrm{A}^*_2]$, where

$$\mathrm{A}^*_1 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

and

$$\mathrm{A}^*_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

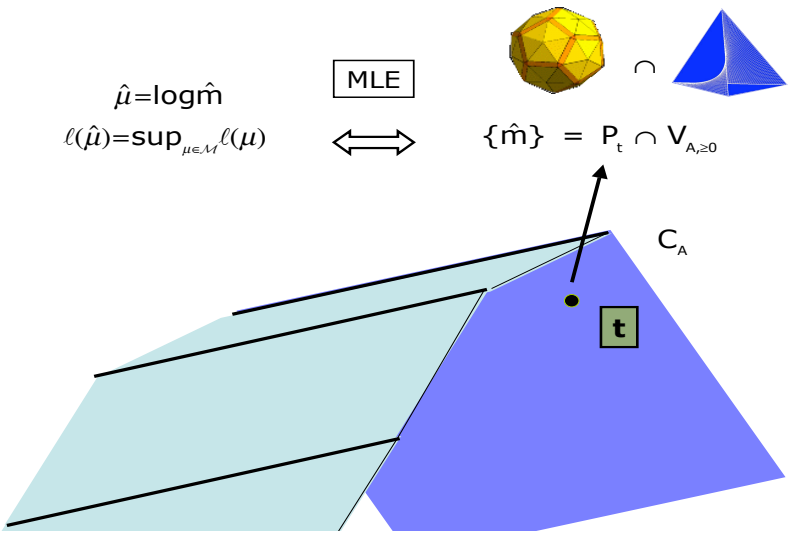
See Fienberg and Rinaldo (2006a) for more details. Then, $\mathrm{kernel}(\mathrm{A})$ contains the set of constant functions in $\mathbb{R}^{\mathcal{I}}$, so that, for any $\boldsymbol{\xi} \in \mathrm{C}_{\mathrm{A}}$, $\mathrm{P}_{\boldsymbol{\xi}}$ contains the ray $\lambda \cdot \mathbf{1}_{\mathcal{I}}$, for any real scalar $\lambda \geq 0$ and is therefore unbounded. As a result, for this models, the MLE of the cell mean vector associated with any table with positive constant entries is always the vector $\mathbf{1}_{\mathcal{I}}$, no matter how large the entries in the tables are. Generally, if $\mathrm{kernel}(\mathrm{A})$ contains vectors of the same sign, the polyhedra $\mathrm{P}_{\boldsymbol{\xi}}$, $\boldsymbol{\xi} \in \mathrm{C}_{\mathrm{A}}$, will be unbounded. For the class of non-hierarchical log-linear models generated by the subspaces of interactions (see Darroch and Speed, 1983), they will be polytopes if and only if the subspace $\mathcal{M}$ contains the constant functions. If the models is hierarchical, this condition is satisfied. ■

The main result of this section is the following theorem which relates toric varieties, marginal cones and the polytopes obtained as the convex hull of the points in the fibers.

Theorem 5.8. $\left\{\mathbf{m}_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi} \in \mathrm{C}_{\mathrm{A}}} = V_{\mathrm{A},\geq 0}$ *and, for each $\boldsymbol{\xi} \in \mathrm{C}_{\mathrm{A}}$, $\mathrm{P}_{\boldsymbol{\xi}} \cap V_{\mathrm{A},\geq 0} = \{\mathbf{m}_{\boldsymbol{\xi}}\} \in \mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}})$.*

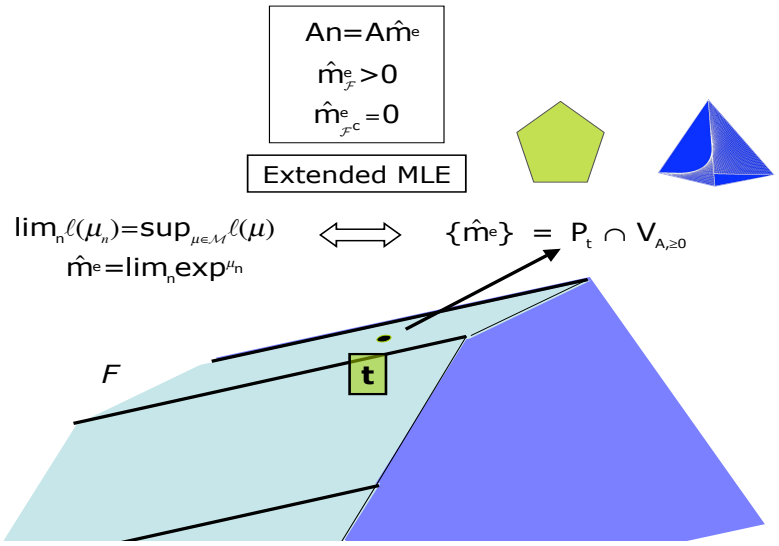
Theorem 5.8 gives a purely geometric interpretation of both the MLE and the extended MLE for log-linear models as the unique points realizing the intersections between the polytopes $\left\{\mathrm{P}_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi}\in\mathrm{C_A}}$ and the variety $V_{\mathrm{A},\geq 0}$. These points happen to be also the optimizers of the log-likelihood functions parametrized by the cell mean vectors for the extended exponential family with convex support $\mathrm{C_A}$. Both the geometric and the analytic characterizations of the extended MLE are illustrated in Figures 5 and 6, where the log-likelihood function, denoted by ℓ, is parametrized for convenience by $\boldsymbol{\mu}$. Figure 5 deals with the MLE, which exists if and only if the vector $\mathbf{t}$ of the observed sufficient statistics lies in the relative interior of the marginal cone $\mathrm{C_A}$. In that case, the intersection between the corresponding polytope $\mathrm{P}_{\mathbf{t}}$, represented as a 3-dimensional polygon, and the variety $V_{\mathrm{A},\geq 0}$ is a unique strictly positive cell mean vector $\widehat{\mathbf{m}}$ whose logarithm is the MLE of $\boldsymbol{\mu}$. The characterization of the extended MLE, presented in Figure 6, is more elaborated but utilizes the same framework. When the vector of observed margins $\mathbf{t}$ lies on the relative interior of a face F of the marginal cone, it is still true that the intersection $V_{\mathrm{A},\geq 0}\cap\mathrm{P}_{\mathbf{t}}$ is realized by a unique non-negative vector $\widehat{\mathbf{m}}^{\mathrm{e}}$. However, since $\mathrm{P}_{\mathbf{t}}$ is not of full dimension (and this is why it is depicted as a 2-dimensional polygon), some of the coordinates of $\widehat{\mathbf{m}}^{\mathrm{e}}$ are 0. Specifically, $\mathrm{supp}(\widehat{\mathbf{m}}^{\mathrm{e}})=\mathcal{F}$, with $\mathcal{F}$ being the facial set for the face F. As in the previous case, the log-likelihood function, parametrized using the logarithm of the cell mean vector, achieves its supremum along sequences of points $\{\boldsymbol{\mu}_n\}$ with exploding norm satisfying $\lim_n \exp\boldsymbol{\mu}_n = \widehat{\mathbf{m}}^{\mathrm{e}}$. This unique limit point $\widehat{\mathbf{m}}^{\mathrm{e}}$ parametrizes, in a mean value fashion, one distribution of to the restricted linear exponential family whose sufficient statistics have F as a convex support, and furthermore, it satisfies the moment equation, $\mathrm{A}\mathbf{n}=\mathrm{A}\widehat{\mathbf{m}}^{\mathrm{e}}$. This point is the extended MLE.

Figure 5: The geometry of the MLE.



Theorem 5.8 also provides a geometric re-interpretation of the generalized iterative propor-

Figure 6: The geometry of the extended MLE.



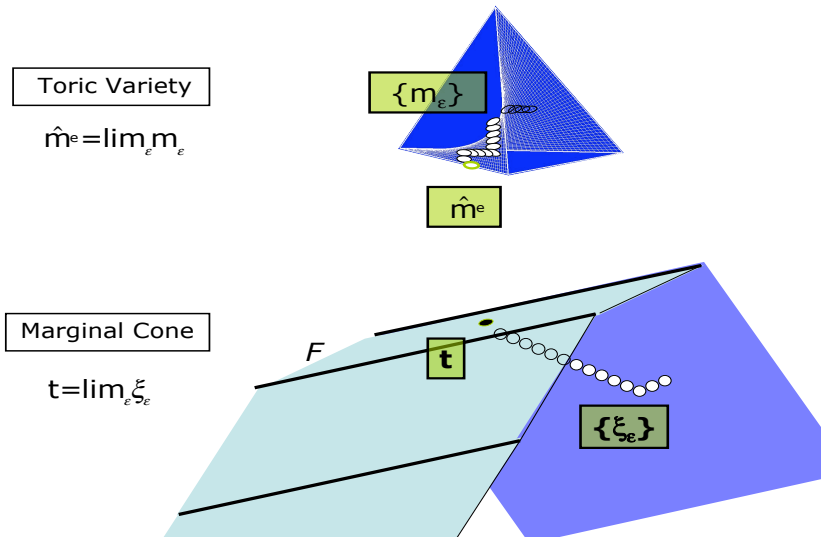
tional fitting (IPF) algorithm (Darroch and Ratcliff, 1972). By performing cyclical adjustments of the fitted values in such a way that the vector of marginal totals tend to the observed margins, the IPF procedure generates a sequence $\{\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\}$ of points in $\mathrm{ri}(\mathrm{C_A})$ such that $\lim_{\boldsymbol{\epsilon}\to\mathbf{0}} \boldsymbol{\xi}_{\boldsymbol{\epsilon}} = \mathbf{t}$, where $\boldsymbol{\xi}_{\boldsymbol{\epsilon}} = \mathbf{t} + \boldsymbol{\epsilon}$, for some vanishing perturbation sequence of values $\boldsymbol{\epsilon}$ like the scalar values in Tables 1**(a)** and 1**(b)**. The sequence$\{\boldsymbol{\xi}_{\boldsymbol{\epsilon}}\}$ is mirrored, in a one-to-one way, by a sequence of strictly positive points $\{\mathbf{m}_{\boldsymbol{\epsilon}}\}$ on the toric variety $V_{\mathrm{A},\geq 0}$ which tend to the extended MLE $\hat{\mathbf{m}}^{\mathrm{e}}$. This correspondence, is represented in Figure 7. Note that, by construction, the IPF algorithm will produce the extended MLE, a results which follows from Theorem 4.13 in Lauritzen (1996), although the convergence has been observed to be extremely slow (see, e.g., Fienberg and Rinaldo, 2006b).

The distribution over the fiber of lattice points inside the polytope $\mathrm{P}_{\mathbf{t}}$ is well known in statistics as the "exact distribution" of the set of all contingency tables possessing the same sufficient statistics $\mathbf{t}$. The study of the geometric and combinatorial properties of $\mathrm{P}_{\mathbf{t}}$ is crucial to many algorithms for sampling from this conditional distributions, in particular to Markov Bases methods (see, for example, Diaconis and Sturmfels, 1998; Takemura and Aoki, 2004; Chen et al., 2006). We remark here that the problem of computing a Markov Bases is in general much harder than determining the existence of the MLE and computing the extended MLE; in fact, if a Markov Basis is available, it can be used to decide whether the MLE exists and to determine the appropriate facial set (Rinaldo, 2005). On the other hand, if one knows the facial set corresponding to the observed sufficient statistics, it is apparent that some Markov moves, namely the ones specified by vectors in the basis with support not inside the facial set, are not applicable and hence need not to be computed.

In general, for conditional Poisson models with convex support $\mathcal{C}_{\mathcal{N}}$, the polyhedra $\mathrm{P}_{\boldsymbol{\xi}}$, $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$, consist each of the sets of all conditional cell mean vectors $\mathbf{m}$ arising from distributions absolutely continuous with respect to $\mu_{\mathcal{N}}$ and satisfying the linear constrains $\mathrm{A}^{*}\mathbf{m} = \boldsymbol{\xi}$. In this respect, they

represented the equivalent of convex cores for the sufficient statistics (see Csiszár and Matúš, 2001, Theorem 3), though $\mathrm{P}_{\boldsymbol{\xi}}$ can be strictly larger than (the closure of) the convex hull of its fiber. Under Poisson and product multinomial schemes, the extended maximum likelihood estimates are the only mean vectors that satisfy both the polynomial equations defining the log-linear model variety and the linear equations implied by the sufficient statistics. In the spirit of (Csiszár and Matúš, 2003, Section VI), the extended MLE can then be interpreted as the rI-projection over the set of all possible distributions over contingency tables having cell mean values satisfying the moment equations.

Figure 7: Homeomorphic correspondence between $V_{\mathrm{A},\geq 0}$ and C_{A} and visual demonstration of the IPF algorithm for computing the extended MLE.



Provided that $\mathcal{M}$ contains the subspace of constant functions, the proof of Theorem 5.8 shows that each point $\mathbf{m}$ in the non-negative toric variety $V_{\mathrm{A},\geq 0}$ is the maximizer of the entropy function over the polytope $\mathrm{P}_{\boldsymbol{\xi}}$, with $\boldsymbol{\xi} = \mathrm{A}\mathbf{m}$. This result should be related to Corollary 4.3, stating the well known result that MLEs for linear exponential families correspond to distributions maximizing Shannon's entropy given the linear constraints associated with the sufficient statistics (see Cover and Thomas, 1991, Section 11.1). Then this result holds also for the extended maximum likelihood estimates.

6 Conclusions

In this article, we make use of various connections between polyhedral geometry, algebraic geometry and statistics to provide analytic and geometric characterizations of maximum-likelihood estimation for log-linear models. Some of these connections are already renown and their usage

is well established among researchers in the field of algebraic statistics. Others, such as the link between toric varieties, marginal cones and extended exponential families, and the consequent geometric representation of both the MLE and the extended MLE, are novel contributions.

We derive new, constructive conditions for the existence of the MLE of the cell mean vector that hold for a variety of sampling schemes, which generalize the results by Haberman (1974) and Eriksson et al. (2005). This new characterization relies on the geometric and combinatorial properties of the marginal cone generated by the design matrix and allows for a unambiguous identification of all the patterns of sampling zeros causing the MLE to be nonexistent. The examples from Section 3.1 exemplify this kind of result.

A series of important results on extended exponential families obtained by Bardorff-Nielsen (1978), Brown (1986) and Csiszár and Matúš (2001, 2003, 2005) are adapted to the conditional distribution of the cell counts under Poisson, product-multinomial and conditional Poisson sampling schemes. In particular, it is shown that these conditional distributions are of exponential types and, therefore, they admit and extended representation. The appropriate mean value parametrization for this extended family is through the point of the toric variety generated by the design matrix. The topological properties of this variety are then used to prove information-theoretic properties of the extended families themselves. From the analytic point of view, we also show that the log-likelihood function of the cell counts always attains its supremum, under any log-linear model and any of the sampling schemes considered here. This supremum corresponds to the usual MLE when the sufficient statistics belong to the relative interior of the marginal cone, while it corresponds to the MLE for a restricted component of the extended family otherwise. This result, combined with the continuity of the mean value parametrization for the extended exponential family allows for a more general definition of the extended MLE than the one proposed by Fienberg et al. (1980) and, only implicitly, by Lauritzen (1996). The extended MLE is unique and always defined and can be computed by maximizing the log-likelihood function of the original, non-extended, family of distributions, without requiring any reparametrization of the likelihood. In fact, numerical procedure for detecting non-existence of the MLE and for computing the extended MLE can be devised, based on the findings presented in this article. The derivation and properties of these algorithms are beyond the scope of this work and are given in Rinaldo (2005) and Fienberg and Rinaldo (2006a).

Extended MLE can be used to correct, in a straightforward way, existing hypothesis testing and model selection procedures to account for the non-estimability of some cell mean vectors due to sampling zeros. On one hand, we already remarked that the knowledge of the facial set determined by the observed table may be of advantage in reducing the computational burden of algorithms for sampling from the fiber. On the other hand, the large-sample χ^2 approximation to various goodness-of-fit statistics is still valid in the extended exponential family framework, provided the appropriate adjustments for the number of degrees of freedom are made (see Fienberg and Rinaldo, 2006a). It is apparent, however, that a careful interpretation of these tests is in order, because they allow for the possibility of "boundary models", entailing cell mean vectors with zero entries.

7 Appendix A

Proof of Proposition 2.1. By definition, the vector $\mathbf{t} = \mathrm{A}\mathbf{n}$ is a minimal sufficient statistic for the exponential family in (3) if and only if the model (2) holds. Using the identity $(\mathbf{t}, \boldsymbol{\eta}) = (\mathbf{n}, \mathrm{A}^\top \boldsymbol{\eta}) =$

$(\mathbf{n}, \boldsymbol{\mu})$, where $\boldsymbol{\mu} = \mathrm{A}^\top \boldsymbol{\eta} \in \mathcal{M}$, this in turn occurs if and only if Equation (5) becomes

$$p_{\boldsymbol{\eta}}(\mathbf{x}) d\nu_{\mathcal{N}}(\mathbf{x}) = \frac{\frac{1}{\prod_i \mathbf{x}(i)!} \exp\{(\mathbf{x}, \boldsymbol{\mu})\}}{\sum_{\mathbf{x}' \in S(\mathcal{N})} \frac{1}{\prod_i \mathbf{x}'(i)!} \exp\{(\mathbf{x}', \boldsymbol{\mu})\}} \frac{d\nu_{\mathcal{N}}}{d\nu}(\mathbf{x}). \tag{23}$$

If $\mathcal{N} = \{\mathbf{0}\}$, so that $S(\mathcal{N}) = \mathbb{N}^{\mathcal{I}}$, (23) is the joint distribution of $|\mathcal{I}|$ independent Poisson random variables with mean vector $\exp \boldsymbol{\mu}$. Thus, $\boldsymbol{\mu} \in \mathcal{M}$ is indeed the logarithm of the unconditional mean vector of the table $\mathbf{n}$. To prove *ii.*, note that, since $\boldsymbol{\mu} \in \mathcal{M}$ and $\mathrm{P}_{\mathcal{M}}$ is a symmetric operator,

$$(\mathbf{n}, \boldsymbol{\mu}) = (\mathrm{P}_{\mathcal{M}} \mathbf{n}, \mathrm{P}_{\mathcal{M}} \boldsymbol{\mu}) = (\mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \mathbf{n}, \mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \boldsymbol{\mu}) + (\mathbf{c}, \mathrm{P}_{\mathcal{N}} \boldsymbol{\mu})$$

Thus, (23) becomes

$$p_{\boldsymbol{\eta}}(\mathbf{x}) d\nu_{\mathcal{N}}(\mathbf{x}) = \frac{\frac{1}{\prod_i \mathbf{x}(i)!} \exp\{(\mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \mathbf{x}, \mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \boldsymbol{\mu})\}}{\sum_{\mathbf{x}' \in S(\mathcal{N})} \frac{1}{\prod_i \mathbf{x}'(i)!} \exp\{(\mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \mathbf{x}', \mathrm{P}_{\mathcal{M} \ominus \mathcal{N}} \boldsymbol{\mu})\}} \frac{d\nu_{\mathcal{N}}}{d\nu}(\mathbf{x}), \tag{24}$$

proving the statement. ■

Proof of Theorem 3.5. The first claim stems from equation (24), from which it follows that, a.e-$\nu_{\mathcal{N}}$,

$$(\mathcal{P}_{\mathcal{M} \ominus \mathcal{N}} \mathbf{n}, \mathcal{P}_{\mathcal{M} \ominus \mathcal{N}} \boldsymbol{\mu}) = (\mathbf{z}, \boldsymbol{\theta}), \tag{25}$$

where $\boldsymbol{\theta} \in \mathbb{R}^k$. Then, the distribution of $\mathbf{z}$ belongs to an exponential family with base measure $\mu_{\mathcal{N}}$. Minimality essentially follows from the sampling assumption **S**, which implies that there are no affine dependencies among the coordinates of $\mathbf{z}$. Similarly, the parameter vector $\boldsymbol{\theta}$ is unconstrained, hence $\boldsymbol{\theta} \in \mathbb{R}^k$, so the family is full and regular. The convex support of this family is $\mathcal{C}_{\mathcal{N}} = \{\mathbf{z} = \mathrm{A}^* \mathbf{x} : \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_+, (\boldsymbol{\gamma}_j, \mathbf{x}) = c_j, j = 1, \ldots, m\}$. Theorem 3.4 shows that the MLE $\widehat{\boldsymbol{\theta}}$ exists and is unique if and only if $\mathbf{z} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$. Next, letting V be the $m \times |\mathcal{I}|$ matrix with rows the vectors $\boldsymbol{\gamma}_j / c_j$, $j = 1, \ldots, m$, $\mathbf{z}$ belongs to $\mathcal{C}_{\mathcal{N}}$ if and only if

$$\begin{pmatrix} \mathbf{z} \\ \mathbf{1} \end{pmatrix} \in \mathrm{cone} \begin{pmatrix} \mathrm{A}^* \\ \mathrm{V} \end{pmatrix}. \tag{26}$$

This can be recognized to be a $\mathcal{V}$-polyhedron (Ziegler, 1998) and its dimension $k = \mathrm{rank}(\mathrm{A}^*)$ is the dimension of its affine hull, which is the order of the exponential family (Brown, 1986). Next, consider the polyhedron $\mathrm{P}_{\mathrm{A}} = \{\mathbf{t} = \mathrm{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0}, (\boldsymbol{\gamma}_j, \mathbf{x}) = c_j, j = 1, \ldots, m\}$. Since $\begin{pmatrix} \mathrm{A}^* \\ \mathrm{V} \end{pmatrix}$ and A have the same null space and both the polyhedra P_{A} and $\mathcal{C}_{\mathcal{N}}$ are specified by the same set of linear forms given by the vectors $\boldsymbol{\gamma}_j$, $j = 1, \ldots, m$, they have the same dimension and Gale transform. As a result, they are combinatorially equivalent, hence $\mathbf{z} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ if and only if $\mathbf{t} \in \mathrm{ri}(\mathrm{P}_{\mathrm{A}})$. By Theorem 3.4, if the MLE $\widehat{\boldsymbol{\theta}}$ exists, it is unique. Since the row range of A^* is $\mathcal{M} \ominus \mathcal{N}$ and A^* is of full rank, the last statement is equivalent to existence and uniqueness of the MLE of $\mathcal{P}_{\mathcal{M} \ominus \mathcal{N}} \boldsymbol{\mu}$, which is $(\mathrm{A}^*)^\top \widehat{\boldsymbol{\theta}}$. ■

Proof of Theorem 3.7. The proof proceeds by reducing to minimal form the corresponding exponential families and then by exploiting the fact that the gradient of the log-partition function at the MLE equals the observed minimal sufficient statistics.

Poisson Sampling
Letting A be the full-column rank design matrix whose columns span $\mathcal{M}$, the distribution of the minimal sufficient statistics belongs to the minimal, full and regular exponential family with density with respect to the induced measure $\nu_{\mathcal{N}} T^{-1}$ given by

$$\exp\{(\mathbf{z}, \boldsymbol{\theta}) - (\mathbf{1}_{\mathcal{I}}, \mathrm{A}^{\top}\boldsymbol{\theta})\}.$$

The log-partition function is $\psi(\boldsymbol{\theta}) = (\mathbf{1}_{\mathcal{I}}, \mathrm{A}^{\top}\boldsymbol{\theta})$ and the natural parameter space is $\mathbb{R}^k$. The MLE $\widehat{\boldsymbol{\theta}}$ exists and is unique if and only if $\mathbf{z} \in \mathrm{ri}(\mathrm{cone}(\mathrm{A}))$ and it satisfies $\nabla\psi(\widehat{\boldsymbol{\theta}}) = \mathrm{A}\exp^{\mathrm{A}^{\top}\widehat{\boldsymbol{\theta}}} = \mathbf{z}$, which implies $\mathrm{A}\widehat{\mathbf{m}} = \mathrm{A}\mathbf{n}$ or, equivalently, $\mathcal{P}_{\mathcal{M}}\widehat{\mathbf{m}} = \mathcal{P}_{\mathcal{M}}\mathbf{n}$.

Product-multinomial Sampling
Recall that the constraint subspace is spanned by the orthogonal vectors $(\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r)$ which are the indicators of the corresponding to a partition of $\mathcal{I}$ in r classes $\mathcal{B}_j = \mathrm{supp}(\boldsymbol{\chi}_j)$, $j = 1, \ldots, r$. Let A be a full-column-rank design matrix spanning $\mathcal{M} \ominus \mathcal{N}$ and let $\mathrm{W} = [\boldsymbol{\chi}_1 \quad \ldots \quad \boldsymbol{\chi}_r]$, so that $\mathcal{R}(\mathrm{W}) = \mathcal{N}$. Then, since $\mathcal{N} \subsetneq \mathcal{M}$, for each $\boldsymbol{\mu} \in \mathcal{M}$, $\boldsymbol{\mu} = \mathrm{A}\boldsymbol{\theta} + \mathrm{W}\boldsymbol{\gamma}$ with $\boldsymbol{\theta} \in \mathbb{R}^k$ and $\boldsymbol{\gamma} \in \mathbb{R}^r$. By construction, any $|\mathcal{I}|$-dimensional vector of the form $\mathrm{W}\boldsymbol{\gamma}$ has constant values γ_j along the coordinates $i \in \mathcal{B}_j$. It is possible to show (see, for example, Haberman, 1974, page 11) that the conditional mean $\mathbf{m}$ satisfies

$$\frac{\mathbf{m}(i)}{N_j} = \frac{\exp^{\mathrm{A}}\boldsymbol{\theta}}{(\exp^{\mathrm{A}}\boldsymbol{\theta}, \boldsymbol{\chi}_j)} \qquad \text{for } i \in \mathcal{B}_j, \tag{27}$$

from which it can be deduced that, for $j = 1, \ldots, r$, $N_j = \exp^{\gamma_j}(\exp^{\mathrm{A}}\boldsymbol{\theta}, \boldsymbol{\chi}_j)$. Therefore, only the vector $\boldsymbol{\theta}$ needs to be estimated since $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\theta})$ can be recovered uniquely from it. Furthermore, since $\boldsymbol{\chi}_j \in \mathcal{M}$, by taking the log of both sides of (27), it follows that $\log \mathbf{m} \in \mathcal{M}$. The exponential family of distribution for the sufficient statistics $\mathbf{z}$ generated by $\nu_{\mathcal{N}}$ and A is minimal, full and regular and has $\mathbb{R}^k$ as its natural parameter space, where $k = \dim(\mathcal{M} \ominus \mathcal{N})$. With respect to the induced measure, the densities take the form

$$\exp\left\{(\mathbf{z}, \boldsymbol{\theta}) - \sum_{j=1}^{r} N_j \log\left((\exp^{\mathrm{A}}\boldsymbol{\theta}, \boldsymbol{\chi}_j)\right)\right\}.$$

The gradient of the log-partition function is

$$\begin{aligned}
\nabla\psi(\boldsymbol{\theta}) &= \textstyle\sum_{j=1}^{r} \frac{N_j}{(\exp^{\mathrm{A}}\boldsymbol{\theta}, \boldsymbol{\chi}_j)} \sum_{i:\, i\in\mathcal{B}_j} \mathbf{a}_i^{\top} \exp^{\mathbf{a}_i^{\top}\boldsymbol{\theta}} \\
&= \textstyle\sum_{j=1}^{r} \exp^{\gamma_j} \sum_{i:\, i\in\mathcal{B}_j} \mathbf{a}_i^{\top} \exp^{\mathbf{a}_i^{\top}\boldsymbol{\theta}} \\
&= \textstyle\sum_{j=1}^{r} \sum_{i:\, i\in\mathcal{B}_j} \mathbf{a}_i^{\top} \exp^{\mathbf{a}_i^{\top}\boldsymbol{\theta} + \mathbf{w}_i^{\top}\boldsymbol{\gamma}} \\
&= \mathrm{A}^{\top} \exp^{\mathrm{A}\boldsymbol{\theta} + \mathrm{W}\boldsymbol{\gamma}},
\end{aligned}$$

where $\mathbf{a}_i^{\top}$ and $\mathbf{w}_i^{\top}$ denote the i-th row of the matrices A and W, respectively. As above, if the MLE $(\widehat{\boldsymbol{\theta}}, \boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}}))$ exists, it is unique and the moment equation $\nabla\psi(\widehat{\boldsymbol{\theta}}) = \mathbf{z}$ holds. This is equivalent to $\mathrm{A}^{\top}\exp^{\mathrm{A}\widehat{\boldsymbol{\theta}} + \mathrm{W}\boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}})} = \mathrm{A}^{\top}\widehat{\mathbf{m}} = \mathrm{A}^{\top}\mathbf{n}$, which in turn implies $\mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\widehat{\mathbf{m}} = \mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\mathbf{n}$. In addition, because of the way the vector $\boldsymbol{\beta}$ is constrained, $\mathcal{P}_{\mathcal{N}}\widehat{\mathbf{m}} = \mathcal{P}_{\mathcal{N}}\mathbf{n}$.

For the second part of the statement, since $\mathcal{N} \subset \mathcal{M}$, the equality $\mathcal{P}_{\mathcal{M}}\widehat{\mathbf{m}} = \mathcal{P}_{\mathcal{M}}\mathbf{n}$ is satisfied for the MLE under both sampling schemes. By the uniqueness, they must exist together and coincide. ■

Proof of Lemma 3.8. Since the columns of $(\mathrm{A}^*)^\top$ span $\mathcal{M} \ominus \mathcal{N}$, $(\mathrm{A}^*)^\top\boldsymbol{\theta} = \mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu}$, for some (in fact, many) $\boldsymbol{\mu} \in \mathcal{M}$. Then, for one $\boldsymbol{\eta} \in \mathbf{H}_{\nu_{\mathcal{N}}}$, $\boldsymbol{\mu} = \mathrm{A}^\top\boldsymbol{\eta}$. From this, it follows that $(\mathrm{A}^*)^\top\boldsymbol{\theta} = \mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\mathrm{A}^\top\boldsymbol{\eta}$, and, because A^* is of full row-rank, $\boldsymbol{\theta} = \mathrm{H}\boldsymbol{\eta}$, where the matrix

$$\begin{aligned} \mathrm{H} &= \left(\mathrm{A}^*(\mathrm{A}^*)^\top\right)^{-1}\mathrm{A}^*\mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\mathrm{A}^\top \\ &= \left(\mathrm{A}^*(\mathrm{A}^*)^\top\right)^{-1}\mathrm{A}^*\mathrm{A}^\top \end{aligned}$$

is independent of $\boldsymbol{\eta}$. Equation (9) follows from writing the denominator of (24) as

$$\int_{S(\mathcal{N})} \exp\{(\mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\mathbf{x}, \mathcal{P}_{\mathcal{M}\ominus\mathcal{N}}\boldsymbol{\mu})\} d\nu_{\mathcal{N}}(\mathbf{x}) = \int_{\mathcal{C}_{\mathcal{N}}} \exp\{(\mathbf{z}, \boldsymbol{\theta})\} d\mu_{\mathcal{N}}(\mathbf{z}) = \exp\{\psi(\boldsymbol{\theta})\},$$

where the equality in the middle stems from the definition of $\mu_{\mathcal{N}}$ and Equation (25). ■

Proof of Lemma 3.9. Using the settings of Theorem 3.5, let $\mathrm{A}' = \begin{pmatrix} \mathrm{A}^* \\ \mathrm{V} \end{pmatrix}$. The matrices A' and A have the same row span, so there exists a matrix T such that $\mathrm{A} = \mathrm{T}\mathrm{A}'$. Let $\mathbf{1}_r$ be a r-dimensional vector of ones, where r is the number of rows of V. For any $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$, let $f_{\mathrm{A}}(\boldsymbol{\xi}) = \mathrm{T}\begin{pmatrix} \boldsymbol{\xi} \\ \mathbf{1}_r \end{pmatrix}$. Then f_{A} maps linearly $\mathcal{C}_{\mathcal{N}}$ into C_{A}. To see that f_{A} is an injection, let $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ be two distinct points of $\mathcal{C}_{\mathcal{N}}$. Then, there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{\mathcal{I}}_{\geq 0}$ such that $\boldsymbol{\xi}_1 = \mathrm{A}^*\mathbf{x}_1$ and $\boldsymbol{\xi}_2 = \mathrm{A}^*\mathbf{x}_2$ and $\mathrm{V}\mathbf{x}_1 = \mathrm{V}\mathbf{x}_2 = \mathbf{1}_r$. Suppose that $f_{\mathrm{A}}(\boldsymbol{\xi}_1) = f_{\mathrm{A}}(\boldsymbol{\xi}_1)$. Then $\mathbf{t}_1 = \mathbf{t}_2$, where

$$\mathbf{t}_i = f_{\mathrm{A}}(\boldsymbol{\xi}_i) = \mathrm{T}\mathrm{A}'\mathbf{x}_i = \mathrm{A}\mathbf{x}_i \in \mathrm{C}_{\mathrm{A}}, \quad i = 1, 2.$$

It follows that $\mathbf{x}_1 - \mathbf{x}_2$ belongs to $\mathrm{kernel}(\mathrm{A})$ and, therefore, to $\mathrm{kernel}(\mathrm{A}^*)$, which implies that $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2$, a contradiction. The final statement can be proved with similar arguments or using Lemma 7.10 in Ziegler (1998) after noticing that f_{A} is the inverse of an affine projection. ■

Proof of Theorem 4.2. This result follows by combining Theorem 9.1 $(ii)^*$ and Theorem 9.4 of Bardorff-Nielsen (1978) with the fact that $\mathcal{C}_{\mathcal{N}}$ is closed.

The alternative proof given here highlights various connections between convex analysis and exponential families. For a minimal and full family as in (8), it is known that $\psi(\boldsymbol{\theta})$ is a strictly convex, closed and essentially smooth function. The conjugate function $\psi^*(\boldsymbol{\xi}) = \{\sup_{\boldsymbol{\theta}\in\Theta}(\boldsymbol{\xi}, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\}$ is also an essentially smooth and closed convex function (Rockafellar, 1970, Theorem 12.2 and Theorem 26.3). For any convex function f from $\mathbb{R}^k$ to $\mathbb{R}$, a vector $\mathbf{x}^* \in \mathbb{R}^k$ is a *subgradient* at a point $\mathbf{x} \in \mathbb{R}^k$ (not necessarily in $\mathrm{dom}(f)$) if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{x}^*, \mathbf{y} - \mathbf{x}) \quad \forall \mathbf{y} \in \mathbb{R}^k.$$

The convex set of all subgradients of f at $\mathbf{x}$ is called *subdifferential* and is denoted with $\partial f(\mathbf{x})$ and the multivalued map $\partial f : \mathbf{x} \to \partial f(\mathbf{x})$ is the subdifferential of f. For a proper convex function such as ψ, $\partial\psi(\mathbf{x}) = \emptyset$ if $\mathbf{x} \notin \mathrm{dom}\psi$ and $\partial\psi(\mathbf{x}) \neq \emptyset$ if $\mathbf{x} \in \mathrm{ri}(\mathrm{dom}\psi)$. By Theorem 23.5 in Rockafellar (1970), since ψ is closed convex function, the following statements are equivalent:

- $\boldsymbol{\xi} \in \partial\psi(\boldsymbol{\theta})$;
- $\boldsymbol{\theta} \in \partial\psi^*(\boldsymbol{\xi})$;
- $(\boldsymbol{\theta}, \boldsymbol{\xi}) - \psi(\boldsymbol{\theta}) = \sup_{\boldsymbol{\zeta}\in\Theta}(\boldsymbol{\zeta}, \boldsymbol{\xi}) - \psi(\boldsymbol{\zeta}) = \psi^*(\boldsymbol{\xi})$.

By Theorem 26.1 in Rockafellar (1970), the subdifferential function $\partial\psi$ is a single-valued mapping and reduces to the gradient function $\nabla\psi$, for $\boldsymbol{\theta} \in \Theta$ and, furthermore, $\partial\psi^* = (\nabla\psi)^{-1}$. It can be deduced that the subdifferential function is, in fact, the mean value parametrization function $\Lambda(\boldsymbol{\theta}) = \partial\psi(\boldsymbol{\theta})$ and that the image of $\partial\psi$ is $\mathrm{ri}(\mathcal{C}_\mathcal{N})$. The above displays then show that the MLE exists if and only if $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_\mathcal{N})$. The pair $(\psi^*, \mathrm{ri}(\mathcal{C}_\mathcal{N}))$ is the *Legendre conjugate* of (ψ, Θ) and $\mathrm{ri}(\mathcal{C}_\mathcal{N}) \subset \mathrm{dom}\psi^*$. By Theorem 26.4 in Rockafellar (1970) the associated Legendre transform is given by

$$\psi^*(\boldsymbol{\xi}) = \left((\nabla\psi)^{-1}(\boldsymbol{\xi}), \boldsymbol{\xi}\right) - \psi\left((\nabla\psi)^{-1}(\boldsymbol{\xi})\right) \tag{28}$$

for $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C})$ and $\boldsymbol{\theta} \in \Theta$.

By Corollary 26.4.1 in Rockafellar (1970) ψ^*, restricted to $\mathrm{ri}(\mathcal{C}_\mathcal{N})$, is the sup-log-likelihood function and it is strictly convex on every subset of $\mathrm{ri}(\mathcal{C}_\mathcal{N})$. In addition, owing to the essential smoothness of ψ^*, for any boundary point $\boldsymbol{\xi} \in \mathrm{bd}(\mathcal{C})$, $\psi^*(\boldsymbol{\xi}) = \lim_n \psi^*(\boldsymbol{\xi}_n)$ for *any* sequence $\{\boldsymbol{\xi}_n\} \subset \mathrm{ri}(\mathcal{C})$ on a line joining $\boldsymbol{\xi}$ with any point in $\mathrm{ri}(\mathcal{C})$. Next, $\mathrm{ri}(\mathrm{dom}\psi^*) \subset \mathrm{ri}(\mathcal{C}_\mathcal{N}) \subset \mathrm{dom}\psi^*$. Therefore, since both $\mathrm{ri}(\mathcal{C}_\mathcal{N})$ and $\mathrm{dom}\psi^*$ are convex sets, $\mathcal{C}_\mathcal{N} = \mathrm{cl}(\mathrm{dom}\psi^*)$. Finally, since ψ^* is closed, $\mathrm{cl}(\mathrm{dom}\psi^*) = \mathrm{dom}\psi^*$ and so conclude $\mathcal{C}_\mathcal{N} = \mathrm{dom}\psi^*$. ■

Proof of Theorem 4.5.

i. The case in which $\mathbf{z} \in \mathrm{ri}(\mathcal{C}_\mathcal{N})$ was already proved and it suffices to show that the MLE exists and is unique for any $\mathbf{z} \in \mathrm{ri}(F)$, for any proper face F of $\mathcal{C}_N$. This proof is based on the arguments utilized by Brown (1986, Theorem 6.21).

Let $\mathbf{e}$ be the normal vector to the supporting hyperplane H_F defining the face F whose relative interior contains $\mathbf{z}$, so that, without loss of generality, $H_F = \{\mathbf{x} : (\mathbf{x}, \mathbf{e}) = 0\}$ and for all $\mathbf{x} \notin F$, $(\mathbf{x}, \mathbf{e}) < 0$. Consider any point $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_\mathcal{N})$ and let $\boldsymbol{\theta} = \Lambda^{-1}(\boldsymbol{\xi})$. Next notice that $(\mathbf{z}, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}) = -\ln \int \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta})} d\mu(\mathbf{x})$. In addition, since $(\mathbf{x} - \mathbf{z}, \mathbf{e}) \leq 0$ for each $\mathbf{x} \in \mathcal{C}_\mathcal{N}$, it follows that $\int \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta}+\rho\mathbf{e})} d\mu(\mathbf{x}) < \int \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta})} d\mu(\mathbf{x})$, for any $\rho > 0$. Therefore, for $\rho \geq 0$,

$$\begin{array}{rcl} \int \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta}+\rho\mathbf{e})} d\mu(\mathbf{x}) & = & \int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z}, \mathbf{e})<0\}} \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta}+\rho\mathbf{e})} d\mu(\mathbf{x}) + \int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z}, \mathbf{e})=0\}} \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta})} d\mu(\mathbf{x}) \\ \downarrow & & \int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z}, \mathbf{e})=0\}} \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta})} d\mu(\mathbf{x}) = \int \exp^{(\mathbf{x}-\mathbf{z}, \boldsymbol{\theta})} d\mu_F(\mathbf{x}), \end{array} \tag{29}$$

as $\rho \to \infty$, by the monotone convergence theorem and because, by construction,

$$\mu_F(B) = \mu(B \cap H_F) = \int_{\mathbf{x}\in B\cap\{\mathbf{x}:(\mathbf{x}-\mathbf{z}, \mathbf{e})=0\}} d\mu(\mathbf{x})$$

for any Borel set B.

Let $\boldsymbol{\xi}' = \Lambda_F(\boldsymbol{\theta})$, which is a unique point in $\mathrm{ri}(F)$, since $\boldsymbol{\theta} \in \Theta = \Theta_F = \mathbb{R}^k$. By taking the negative log of both side in the last display, it follows that

$$p_{\boldsymbol{\xi}}(\mathbf{z}) = \frac{\exp^{(\boldsymbol{\theta}, \mathbf{z})}}{\int \exp^{(\boldsymbol{\theta}, \mathbf{x})} d(\mathbf{x})} < \frac{\exp^{(\boldsymbol{\theta}, \mathbf{z})}}{\int \exp^{(\boldsymbol{\theta}, \mathbf{x})} d\mu_{F_{\boldsymbol{\xi}'}}(\mathbf{x})} \frac{d\mu_F}{d\mu}(\mathbf{z}) = p_{\boldsymbol{\xi}'}(\mathbf{z}).$$

Since, for any $\boldsymbol{\xi}'' \in \mathrm{bd}(\mathcal{C}_{\mathcal{N}}) \cap F^c$, $p_{\boldsymbol{\xi}''}(\mathbf{z}) = 0$, this shows that for any $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}} \cap F^c$, there is a $\boldsymbol{\xi}' \in \mathrm{ri}(F)$ such that $p_{\boldsymbol{\xi}}(\mathbf{z}) < p_{\boldsymbol{\xi}'}(\mathbf{z})$. Finally, by Theorem 3.4, applied to the family whose convex support is F, the MLE exists uniquely and coincides with $\mathbf{z}$, so that, for each $\boldsymbol{\xi}' \in F$ with $\boldsymbol{\xi}' \neq \mathbf{z}$, $p_{\mathbf{z}}(\mathbf{z}) > p_{\boldsymbol{\xi}'}(\mathbf{z})$. Note that, in order to apply Theorem 3.4, one has to reduce the exponential family with convex support F to its minimal form; however, it is easy to see that this can be achieved by embedding F into its support hyperplane.

ii. This follows from Section 6.18-6.23 of Brown (1986) and because the convex support is a polyhedron, $\mu_F > 0$ for each face F and the original exponential family if full. For a simpler proof, let $\{\boldsymbol{\xi}_n\} \subset \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ be a sequence such that $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi}$ with $\boldsymbol{\xi} \in \mathrm{ri}(F)$ for some proper face F. By Lemma 7.2, there exists a sequence $\{\boldsymbol{\eta}_n\}$ with $\boldsymbol{\eta}_n = \Lambda^{-1}(\boldsymbol{\xi}_n)$ such that $\lim_n(\mathbf{x}, \boldsymbol{\eta}_n) = (\mathbf{x}, \boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Lambda_F^{-1}(\boldsymbol{\xi})$, for every $\mathbf{x} \in F$ and $\lim_n(\mathbf{x}, \boldsymbol{\eta}_i) = -\infty$ for every $\mathbf{x} \in F^c$. Let $\mathbf{z}$ be any point in $\mathrm{supp}(\mu_{\mathcal{N}})$. Then, using the monotone convergence theorem,

$$\begin{aligned}
\lim_n p_{\boldsymbol{\xi}_n}(\mathbf{z}) &= \lim_n \left(\int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z},\mathbf{e})<0\}} \exp^{(\mathbf{x}-\mathbf{z},\boldsymbol{\eta}_n)} d\mu(\mathbf{x}) + \int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z},\mathbf{e})=0\}} \exp^{(\mathbf{x}-\mathbf{z},\boldsymbol{\eta}_n)} d\mu(\mathbf{x}) \right) \\
&= \int_{\{\mathbf{x}:(\mathbf{x}-\mathbf{z},\mathbf{e})=0\}} \exp^{(\mathbf{x}-\mathbf{z},\boldsymbol{\theta})} d\mu(\mathbf{x}) \\
&= \int \exp^{(\mathbf{x}-\mathbf{z},\boldsymbol{\theta})} d\mu_F(\mathbf{x}) \\
&= p_{\boldsymbol{\xi}}(\mathbf{z}).
\end{aligned}$$

iii. For any $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$, let $\xi^*(\boldsymbol{\xi}) = \sup_{\boldsymbol{\xi}' \in \mathcal{C}_{\mathcal{N}}} p_{\boldsymbol{\xi}'}(\boldsymbol{\xi})$. Using part **i.**, one sees that $\xi^*(\boldsymbol{\xi}) = p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$. In particular, for $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$, $p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \exp\left((\boldsymbol{\xi}, \Lambda^{-1}(\boldsymbol{\xi})) - \psi(\Lambda^{-1}(\boldsymbol{\xi}))\right)$, by the properties of the mean parametrization map. It follows that $(\mathrm{ri}(\mathcal{C}_{\mathcal{N}}), \xi^*)$ is the Legendre conjugate of (Θ, ψ) so that, for $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$,

$$\xi^*(\boldsymbol{\xi}) = \sup_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{\theta}}(\boldsymbol{\xi}) = \psi^*(\boldsymbol{\theta})$$

(see Rockafellar, 1970, page 256-257). Therefore, $\psi^* = \xi^*$, on $\mathrm{ri}(\mathcal{C}_{\mathcal{N}})$. It only remains to show that ψ^* and ξ^* agree also on the boundary of $\mathcal{C}_{\mathcal{N}}$. Let $\{\boldsymbol{\xi}_n\} \subset \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ be an arbitrary sequence such that $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi} \in \mathrm{bd}(\mathcal{C}_{\mathcal{N}})$. By part **ii.**, for each $\epsilon > 0$, there exists a number $N_1(\epsilon)$ such that $\forall n > N_1(\epsilon)$, $|p_{\boldsymbol{\xi}_n}(\boldsymbol{\xi}) - p_{\boldsymbol{\xi}}(\boldsymbol{\xi})| < \epsilon/2$. Next, using part **ii.** again and because of the continuity of the inner product, there exists a number $N_2(\epsilon)$ such that $\forall n > N_2(\epsilon)$, $|p_{\boldsymbol{\xi}_n}(\boldsymbol{\xi}_n) - p_{\boldsymbol{\xi}_n}(\boldsymbol{\xi})| < \epsilon/2$. Let $N(\epsilon) = \max(N_1(\epsilon), N_2(\epsilon))$. Then, for all $n > N(\epsilon)$, $|p_{\boldsymbol{\xi}_n}(\boldsymbol{\xi}_n) - p_{\mathbf{z}}(\mathbf{z})| < \epsilon$. Hence,

$$\xi^*(\boldsymbol{\xi}) = \lim_n \xi^*(\boldsymbol{\xi}_n) = \lim_n \psi^*(\boldsymbol{\xi}_n) = \psi^*(\boldsymbol{\xi}),$$

where the last equality follows from the continuity of ψ^* (see Theorem 4.2).

■

Proof of Lemma 4.6. Since F is a face of $\mathcal{C}_{\mathcal{N}}$, there exist some some vector $\boldsymbol{\zeta}_F \in \mathbb{R}^k$ and scalar c_F defining the hyperplane $H = \{\mathbf{z} \in \mathbb{R}^k : (\mathbf{z}, \boldsymbol{\zeta}_F) = c_F\}$, such that $F = \mathcal{C}_{\mathcal{N}} \cap H$ and $(\boldsymbol{\xi}, \boldsymbol{\zeta}_F) < c_F$ for all $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}} \cap F^c$. For each $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$ consider the polyhedron

$$\mathrm{P}_{\boldsymbol{\xi}} := \{\mathbf{x} \in \mathbb{R}^{\mathcal{I}}_{\geq 0} \colon \mathrm{A}^*\mathbf{x} = \boldsymbol{\xi}, (\mathbf{x}, \boldsymbol{\gamma}_j) = c_j, j = 1, \ldots, m\},$$

where the integer vectors $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)$ span $\mathcal{N}$ and the rows of the integer matrix A^* span $\mathcal{M} \ominus \mathcal{N}$. Then, $\boldsymbol{\xi} \in F$ if and only if $c_F = (\boldsymbol{\xi}, \boldsymbol{\zeta}_F) = (\mathbf{x}, \boldsymbol{\zeta}_F^\top \mathrm{A}^*) = (\mathbf{x}, \boldsymbol{\gamma}_F)$, for all $\mathbf{x} \in \mathrm{P}_{\boldsymbol{\xi}}$, with $\boldsymbol{\gamma}_F = \boldsymbol{\zeta}_F^\top \mathrm{A}^* \in$

$\mathcal{M} \ominus \mathcal{N}$. Equation (26) shows that $\mathcal{C}_{\mathcal{N}}$ is the coordinate projection of a polyhedron which is the Minkowski sum of a polytope and a cone (with one of them being possible empty). Therefore, if F is a face of $\mathcal{C}_{\mathcal{N}}$, then it derives either from a face of the cone or from a face of the polytope. Assume without loss of generality that F correspond to the face of the cone (the proof for a face arising form the polytope is similar and requires just minor modifications involving an homogeneization argument), so that the hyperplane equation is $H = \{\mathbf{z} : (\mathbf{z}, \boldsymbol{\zeta}_F) = 0\}$. Then, given any $\boldsymbol{\xi} \in F$,

$$(\mathbf{x}, \boldsymbol{\zeta}_F^\top \mathrm{A}^*) = 0 \quad \text{for every } \mathbf{x} \in \mathrm{P}_{\boldsymbol{\xi}}. \tag{30}$$

Let $\mathcal{F} := \{i \in \mathcal{I} : (\boldsymbol{\zeta}_F, \mathbf{a}_i) = 0\}$, where $\mathbf{a}_i$ denotes the i-th column of A^*. Since, for each $i \notin \mathcal{F}$, $(\boldsymbol{\zeta}_F, \mathbf{a}_i) < 0$, equation (30) implies that, if $\boldsymbol{\xi} \in F$, then, necessarily, $\mathbf{x}(i) = 0$ for all $\mathbf{x} \in \mathrm{P}_{\boldsymbol{\xi}}$ and all $i \notin \mathcal{F}$. By applying the same arguments as in Theorem 3.5, it can be concluded that the face F is a polyhedron which corresponds to the convex support for the minimal sufficient statistic for the conditional Poisson model with subspace constraint $\mathcal{N}_F$ given by the span of the vectors $\boldsymbol{\gamma}_j, j = 1, \ldots, m$ and the vector $\boldsymbol{\gamma}_F$. In particular, since $\boldsymbol{\gamma}_F \in \mathcal{M} \ominus \mathcal{N}$, then $\mathcal{N} \subset \mathcal{N}_F \subset \mathcal{M}$. Therefore, each table $\mathbf{n}$ arising from such extended sampling schemes must necessarily satisfy $\mathbf{n}(i) = 0$, for each $i \notin \mathcal{I}$. Furthermore, since $\boldsymbol{\xi}$ belongs to the relative interior of only one face F, the uniqueness of $\mathcal{F}$ follows.

■

Proof of Corollary 4.7. Lemma 4.6 shows that for any $\mathbf{x} \in \mathrm{P}_{\boldsymbol{\xi}}$, with $\boldsymbol{\xi} \in \mathrm{ri}(F)$ for some face of $\mathcal{C}_{\mathcal{N}}$ with facial set $\mathcal{F}$, $\mathrm{supp}(\mathbf{x}) \subseteq \mathcal{F}$. The MLE does not exist if and only if $\mathrm{A}^*\mathbf{n} = \mathbf{z} \in \mathrm{ri}(F)$ for some proper face F of $\mathcal{C}_{\mathcal{N}}$, from which the statement follows. ■

Proof of Corollary 4.8. We only need to show that $\pi_{\mathcal{F}}(\mathcal{N}) = \pi_{\mathcal{F}}(\mathcal{N}_F)$, for all non-trivial facial sets $\mathcal{F}$ of $\mathcal{C}_{\mathcal{N}}$. Let $\mathcal{F}$ be any such a set. Then, there exists a vector $\boldsymbol{\zeta}_F$ such that $\mathcal{F} = \{i \in \mathcal{I} \colon (\mathbf{a}_i, \boldsymbol{\zeta}_F) = 0\}$, with $\mathbf{a}_i$ denoting the i-th column of A. Then $\mathcal{N}_F = \mathcal{N} + \mathcal{R}(\boldsymbol{\gamma}_F)$, where $\boldsymbol{\gamma}_F = \boldsymbol{\zeta}_F^\top \mathrm{A}$ and $\mathcal{R}(\boldsymbol{\gamma}_F)$ denotes the one-dimensional subspace spanned by $\boldsymbol{\gamma}_F$. Since $\pi_{\mathcal{F}}(\boldsymbol{\gamma}_F)) = \pi_{\mathcal{F}}(\mathcal{R}(\boldsymbol{\gamma}_F)) = \mathbf{0}$, the claim follows. ■

Proof of Theorem 4.9. Consider a point $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$. Then, the following diagram illustrates bijections that are all homeomorphisms,

$$\begin{array}{ccc} \boldsymbol{\theta} & \Longleftrightarrow & \boldsymbol{\mu} \\ \Updownarrow & & \Updownarrow \\ \boldsymbol{\xi} & & \mathbf{m}_{\boldsymbol{\xi}}, \end{array}$$

with $\mathbf{m}_{\boldsymbol{\xi}}$ satisfying Equation (14). Hence $\{\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})\}$ is homeomorphic to $\{\mathbf{m}_{\boldsymbol{\xi}}, \boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})\}$. The same kind of results holds for the relative interior of any proper face F of $\mathcal{C}_{\mathcal{N}}$ (previous reduction to a minimal representation). More precisely, for any proper face F of the convex support, there is a homeomorphism

$$\boldsymbol{\xi} \Longleftrightarrow \pi_F(\boldsymbol{\mu}),$$

where $\boldsymbol{\xi} \in \mathrm{ri}(F)$, and $\mathbf{m}_{\boldsymbol{\xi}} = \tau_F(\exp^{\pi_F(\boldsymbol{\mu})})$ satisfies Equation (14). Therefore, $\mathrm{ri}(F)$ and $\{\mathbf{m}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \mathrm{ri}(F)}$ are homeomorphic for every F in the face lattice $L(\mathcal{C}_{\mathcal{N}})$. Next, if $\lim_n \mathbf{m}_{\boldsymbol{\xi}_n} = \mathbf{m}_{\boldsymbol{\xi}}$ with $\{\boldsymbol{\xi}_n\} \subset \mathcal{C}_{\mathcal{N}}$ and $\boldsymbol{\xi} \in \mathrm{ri}(F)$ for some proper face F, it is clear that $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi}$. The converse is also true by Lemma 7.1 below. Hence, by the closeness of $\mathcal{C}_{\mathcal{N}}$, the result follows.

■

Lemma 7.1. *Let $\{\boldsymbol{\xi}_n\} \subset \mathcal{C}_{\mathcal{N}}$ be such that $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi}$. Then, $\lim_n \mathbf{m}_{\boldsymbol{\xi}_n} = \mathbf{m}_{\boldsymbol{\xi}}$.*

Proof. The proof will only be given for the case $\boldsymbol{\xi} \in \mathrm{ri}(F)$, for some proper face F, the case $\boldsymbol{\xi} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ being analogous and simpler (in fact $\mathrm{supp}(\mathbf{m}_{\boldsymbol{\xi}}) = \mathcal{I}$). It will be shown that to each sequence $\{\boldsymbol{\xi}_n\} \subset \mathcal{C}_{\mathcal{N}}$, there corresponds a sequence of real valued vectors $\{\boldsymbol{\mu}_n\}$ in $\mathcal{M}$ such that $\lim_n \pi_{\mathcal{F}}(\boldsymbol{\mu}_n) = \boldsymbol{\mu}_{\mathcal{F}}$, with $\boldsymbol{\mu}_{\mathcal{F}} = \exp\{\pi_{\mathcal{F}}(\mathbf{m}_{\boldsymbol{\xi}})\}$ and $\lim_n \pi_{\mathcal{F}^c}(\boldsymbol{\mu}_n) \downarrow -\infty$, element-wise, where $\mathcal{F}^c = \mathcal{I} \setminus \mathcal{F}$. From this it follows that $\lim_n \mathbf{m}_{\boldsymbol{\xi}_n} = \lim_n \exp\{\boldsymbol{\mu}_n\} = \mathbf{m}_{\boldsymbol{\xi}}$.

Since $\boldsymbol{\xi} \in \mathrm{ri}(F)$, there exists a vector (in fact, many) $\boldsymbol{\theta} \in \mathbb{R}^k$ such that $\pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\theta}) = \boldsymbol{\mu}_{\mathcal{F}}$, where $\boldsymbol{\xi} = \mathrm{A}^* \tau_{\mathcal{F}}\left(\exp \boldsymbol{\mu}_{\mathcal{F}}\right)$. By Lemma 7.2, it can be assumed that $\boldsymbol{\xi}_n = \Lambda^{-1}(\boldsymbol{\gamma}_n + \rho_n \mathbf{e})$ where $\boldsymbol{\gamma}_n \to \boldsymbol{\gamma}$, with $(\boldsymbol{\gamma}, \mathbf{x}) = (\boldsymbol{\theta}, \mathbf{x}) + K$ a.e.-μ_F for some constant K, $\mathbf{e}$ is a vector normal to the hyperplane containing F and $0 < \rho_n \to \infty$. Then, $\pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\theta}) = \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\gamma})$. Next note that $\boldsymbol{\mu}_n = (\mathrm{A}^*)^\top(\boldsymbol{\gamma}_n + \rho_n \mathbf{e})$ and hence $\boldsymbol{\mu}_n \in \mathcal{M}$ for each n. Because $\mathbf{e}$ is normal to F, $\pi_{\mathcal{F}}((\mathrm{A}^*)^\top \mathbf{e}) = \mathbf{0}$, and therefore $\pi_{\mathcal{F}}(\boldsymbol{\mu}_n) = \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\gamma}_n)$. Moreover, $\pi_{\mathcal{F}^c}(\boldsymbol{\mu}_n) \downarrow -\infty$ element-wise, because $\pi_{\mathcal{F}^c}((\mathrm{A}^*)^\top \mathbf{e}) < \mathbf{0}$, $\rho_n \uparrow \infty$ and $\boldsymbol{\gamma}_n$ is eventually bounded in each coordinate since it converges to the finite vector $\boldsymbol{\gamma}$. By the continuity of the function $\pi_{\mathcal{F}}$, it follows that

$$\begin{aligned}
\lim_n \pi_{\mathcal{F}}(\boldsymbol{\mu}_n) &= \lim_n \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\gamma}_n) \\
&= \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \lim_n \boldsymbol{\gamma}_n) \\
&= \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\gamma}) \\
&= \pi_{\mathcal{F}}((\mathrm{A}^*)^\top \boldsymbol{\theta}) \\
&= \boldsymbol{\mu}_{\mathcal{F}}.
\end{aligned}$$

■

Lemma 7.2. *Let F be a face of the convex support $\mathcal{C}_{\mathcal{N}}$ and let $\mathbf{e}$ be a normal vector to the hyperplane containing F, so that for any $\mathbf{x} \subset \mathcal{C}_{\mathcal{N}}$, $(\mathbf{x}, \mathbf{e}) \leq c_F$ with equality if and only if $\mathbf{x} \in F$. For any $\boldsymbol{\xi}_F \in \mathrm{ri}(F)$, let $\{\boldsymbol{\xi}_n\} \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ be a sequence such that $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi}_F$. Let $\boldsymbol{\eta}_n = \Lambda^{-1}(\boldsymbol{\xi}_n)$, for each n. Then, $\{\boldsymbol{\eta}_n\}$ can be written as $\boldsymbol{\eta}_n = \boldsymbol{\gamma}_n + \rho_n \mathbf{e}$, with $\lim_n \boldsymbol{\gamma}_n = \boldsymbol{\gamma} \in \Lambda_F^{-1}(\boldsymbol{\xi}_F)$ and $0 < \rho_{i_j} \to \infty$.*

Proof. Assume, without loss of generality that $c_F = 0$. The hypothesis $\lim_n \boldsymbol{\xi}_n = \boldsymbol{\xi}_F$ implies that, for some $\boldsymbol{\theta} \in \Lambda_F^{-1}(\boldsymbol{\xi}_F)$,

$$\lim_n \frac{\int_{\mathbf{x}:\,(\mathbf{x},\mathbf{e})<0} \mathbf{x} \exp^{(\mathbf{x},\boldsymbol{\eta}_n)} d\mu(\mathbf{x}) + \int_{\mathbf{x}:\,(\mathbf{x},\mathbf{e})=0} \mathbf{x} \exp^{(\mathbf{x},\boldsymbol{\eta}_n)} d\mu(\mathbf{x})}{\int_{\mathbf{x}:\,(\mathbf{x},\mathbf{e})<0} \exp^{(\mathbf{x},\boldsymbol{\eta}_n)} d\mu(\mathbf{x}) + \int_{\mathbf{x}:\,(\mathbf{x},\mathbf{e})=0} \exp^{(\mathbf{x},\boldsymbol{\eta}_n)} d\mu(\mathbf{x})} = \frac{\int \mathbf{x} \exp^{(\mathbf{x},\boldsymbol{\theta})} d\mu_F(\mathbf{x})}{\int \exp^{(\mathbf{x},\boldsymbol{\theta})} d\mu_F(\mathbf{x})}.$$

From this identity, it follows that, for each $\mathbf{x} \in \mathrm{supp}(\mu_F)$, $\lim_n(\mathbf{x}, \boldsymbol{\eta}_n) = (\mathbf{x}, \boldsymbol{\gamma})$, with $(\mathbf{x}, \boldsymbol{\gamma}) = (\mathbf{x}, \boldsymbol{\theta})$ a.e.-μ_F. Hence $\boldsymbol{\gamma} \in \Lambda_F^{-1}(\boldsymbol{\xi}_F)$ and, for each $\mathbf{x} \in F^c \cap \mathrm{supp}(\mu)$, $\lim_n(\mathbf{x}, \boldsymbol{\eta}_n) \to -\infty$. Since $\mathcal{C}_{\mathcal{N}}$ is the convex hull of the points in the support of $\mu_{\mathcal{N}}$, this in turn implies that for each $\boldsymbol{\xi} \in F$, $\lim_n(\boldsymbol{\xi}, \boldsymbol{\eta}_n) = (\boldsymbol{\xi}, \boldsymbol{\gamma}) = (\boldsymbol{\xi}, \boldsymbol{\theta})$, while for each $\boldsymbol{\xi} \in F^c \cap \mathcal{C}_{\mathcal{N}}$, $\lim_n(\boldsymbol{\xi}, \boldsymbol{\eta}_n) \to -\infty$. Then, the sequence $\{\boldsymbol{\eta}_n\}$ can be chosen of the form $\boldsymbol{\eta}_n = \boldsymbol{\gamma}_n + \rho_n \mathbf{e}$ with $\lim_n \boldsymbol{\gamma}_n \in \Lambda_F^{-1}(\boldsymbol{\xi}_F)$. ■

Proof of Theorem 4.10. For any $\mathbf{n}$ such that $\mathbf{z} = \mathrm{A}^* \mathbf{n}$, $p_{\boldsymbol{\xi}}(\mathbf{z}) = p_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{n})$. Combine this with Theorem 4.5 and Theorem 4.9 to obtained the desired result. Note that the moment equations are satisfied because

$$\mathbf{z} = \nabla \psi_F(\boldsymbol{\theta}_{\mathbf{z}}) = \mathrm{A}_{\mathcal{F}} \widehat{\mathbf{m}},$$

with $\widehat{\mathbf{m}}$ being the MLE of the conditional mean, where the first equality stems from Theorem 3.4 and the second from the fact that the first moment can be obtained by differentiating the log-partition function. ■

Proof of Theorem 4.12. A proof can be given based on general results in Csiszár and Matúš (2003, 2005). We give here instead a direct proof. For any $\boldsymbol{\xi} \in \mathcal{C}_{\mathcal{N}}$, let $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$ be the distribution corresponding to the density $p_{\mathbf{m}_{\boldsymbol{\xi}}}$ from (12) parametrized by $\mathbf{m}_{\boldsymbol{\xi}}$. Because $V_{\mathrm{A},\geq 0}$ is a closed set in $\mathbb{R}^{\mathcal{I}}$, it is sufficient to prove the following claim. Let $\xrightarrow{\text{tv}}$ and $\xrightarrow{\text{rI}}$ denote convergence in total variation and rI-convergence, respectively.

For any sequence $\{\mathbf{m}_{\boldsymbol{\xi}_n}\} \subset V_{\mathrm{A},\geq 0}$ such that $\mathbf{m}_{\boldsymbol{\xi}_n} \to \mathbf{m}_{\boldsymbol{\xi}}$, $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}} \xrightarrow{\text{rI}} \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$ and $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}} \xrightarrow{\text{tv}} \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$.

We first show *rI*-convergence. Consider any $\mathbf{m}_{\boldsymbol{\xi}}$ and let $\mathrm{A}^* \mathbf{m}_{\boldsymbol{\xi}} = \boldsymbol{\xi} \in \mathrm{ri}(F)$ for some proper face F of $\mathcal{C}_{\mathcal{N}}$, so that $\mathrm{supp}(\mathbf{m}_{\boldsymbol{\xi}}) = \mathcal{F}$. Although $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$ it is defined over $\mathbb{R}^{\mathcal{I}}$, $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}\left(\pi_{\mathcal{F}^c}(\mathbb{R}^{\mathcal{I}}) = \mathbf{0}\right) = 1$. Let $\{\mathbf{m}_{\boldsymbol{\xi}_n}\}$ be any sequence such that $\mathbf{m}_{\boldsymbol{\xi}_n} \to \mathbf{m}_{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}_n \in \mathrm{ri}(\mathcal{C}_{\mathcal{N}})$ for all i. This implies that, for all i, $\mathrm{supp}\left(\mathbf{m}_{\boldsymbol{\xi}_n}\right) = \mathcal{I}$ and, furthermore, $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}} \ll \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}}$. Thus, $D\left(\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}} || \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}}\right)$ is finite and, using the function $\tau_{\mathcal{F}}$ defined in (13), equal to

$$\int_{\mathbb{R}^{\mathcal{I}}} \ln \frac{\exp\{(\mathbf{x}, \tau_{\mathcal{F}}(\boldsymbol{\mu})) - \psi_F(\boldsymbol{\theta})\}}{\exp\{(\mathbf{x}, \boldsymbol{\mu}_n) - \psi(\boldsymbol{\theta}_n)\}} d\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{x}), \tag{31}$$

where $\boldsymbol{\theta} \in \Lambda_F^{-1}(\boldsymbol{\xi})$, $\boldsymbol{\theta}_n = \Lambda^{-1}(\boldsymbol{\xi}_n)$, $\boldsymbol{\mu}_n = \ln \mathbf{m}_{\boldsymbol{\xi}_n}$ and $\boldsymbol{\mu} = \ln \pi_{\mathcal{F}}(\mathbf{m}_{\boldsymbol{\xi}})$. Equation (31) in turn is equal to

$$\int_{\mathbb{R}^{\mathcal{I}}} \left(\tau_{\mathcal{F}}(\pi_{\mathcal{F}}(\mathbf{x})), \tau_{\mathcal{F}}(\boldsymbol{\mu} - \pi_{\mathcal{F}}(\boldsymbol{\mu}_n))\right) d\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}(\mathbf{x}) + \psi(\boldsymbol{\theta}_n) - \psi_F(\boldsymbol{\theta}).$$

By the same arguments used in the proof of Theorem 4.5, $\psi(\boldsymbol{\theta}_n) \to \psi_F(\boldsymbol{\theta})$ and by Lemma 7.1 $\pi_{\mathcal{F}}(\boldsymbol{\mu}_n) \to \boldsymbol{\mu}$. Hence $(\mathbf{x}, \tau_{\mathcal{F}}(\boldsymbol{\mu} - \pi_{\mathcal{F}}(\boldsymbol{\mu}_n))) \to 0$ a.e.-$\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$. Thus by the dominated convergence theorem, the integral vanishes, showing that $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}} \xrightarrow{\text{rI}} \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$.

Convergence in total variation follows from Scheffe's Theorem, after noting that $\nu_{\mathcal{N}}$ is a finite measure. Alternatively, by Pinsker's inequality (Cover and Thomas, 1991, Lemma 12.6.1)

$$\frac{1}{2 \ln 2} \left(||\mathbb{P} - \mathbb{Q}||_1\right)^2 \leq D(\mathbb{P}||\mathbb{Q}),$$

where $||\cdot||_1$ denotes the L_1 norm, it can be seen that rI-convergence implies convergence in total variation. Since we just established rI-convergence, we have that $\mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}_n}} \xrightarrow{\text{tv}} \mathbb{P}_{\mathbf{m}_{\boldsymbol{\xi}}}$.

■

Proof Lemma 5.1. The identity in *i)* follows by taking $\mathbf{z} = \exp^{\boldsymbol{\eta}}$, so that $\mathbf{x} \in \mathrm{im}(\phi_A)$ if and only if $\log(\mathbf{x}) = \mathrm{A}^\top \boldsymbol{\eta} \in \mathcal{M}$. For the statement in *ii)*, note that, $p_{\boldsymbol{\eta}}$ satisfies (2) if and only if $\log p_{\boldsymbol{\eta}} = \mathrm{A}^\top \boldsymbol{\eta} + k_{\boldsymbol{\eta}} \mathbf{1}_{\mathcal{I}}$ for some normalizing constant $k_{\boldsymbol{\eta}}$ depending on $\phi(\boldsymbol{\eta})$. If the row span of A contains the vector $\mathbf{1}_{\mathcal{I}}$ then it is immediate to see that $\log p_{\boldsymbol{\eta}} \in \mathcal{M}$ so $c_{\boldsymbol{\eta}}$ is 1. If this is not the case, the result still holds with $c_{\boldsymbol{\eta}} = \exp^{-k\boldsymbol{\eta}}$. ■

Proof of Lemma 5.6. For the design matrix A, a set $\mathcal{F} \subseteq \mathcal{I}$ is facial if and only if there exists a face F in the cone generated by the columns of A whose relative interior is spanned by conic combinations of the columns of A with positive coefficients along the coordinates in $\mathcal{F}$ and zero

coefficients along all the others. This occurs if and only if for each $\mathbf{u} \in \text{kernel}(\mathrm{A})$, either both $\text{supp}(\mathbf{u}^+)$ and $\text{supp}(\mathbf{u}^-)$ are subsets of $\mathcal{F}$ of neither of them is. Next since $\mathcal{L}_\mathrm{A}$ spans the subspace $\text{kernel}(\mathrm{A})$, it is sufficient to consider only vectors $\mathbf{u} \in \mathcal{L}_\mathrm{A}$. Since each $\mathbf{m} \in V_{\mathrm{A},\geq 0}$ satisfies the polynomial equations (18) defining the toric ideal $\mathcal{I}_\mathrm{A}$, then, for each $\mathbf{u} \in \mathcal{L}_\mathrm{A}$, either both $\text{supp}(\mathbf{u}^+)$ and $\text{supp}(\mathbf{u}^-)$ are subsets of $\text{supp}(\mathbf{m})$ of neither of them is (in which case the equality 0=0 trivially holds). Hence, facial sets and the support sets of point in $V_{\mathrm{A},\geq 0}$ are determined by the same conditions. See also Lemma 4 in Geiger et al. (2006). ■

Proof of Theorem 5.8. The second statements is proved in Lemma 7.3, so only the first claim needs to be proved. By Lemma 7.3, C_A and $V_{\mathrm{A},\geq 0}$ are in one-to-one correspondence, while Theorem 4.9 shows that C_A and $\left\{\mathbf{m}_{\boldsymbol{\xi}}\right\}_{\boldsymbol{\xi}\in \mathrm{C}_\mathrm{A}}$ are homeomorphic. Therefore, it remains to show that, for each $\boldsymbol{\xi} \in \mathrm{C}_\mathrm{A}$, the corresponding point $\mathbf{m}$ in $V_{\mathrm{A},\geq 0}$ is in fact $\mathbf{m}_{\boldsymbol{\xi}}$. If $\boldsymbol{\xi} \in \text{ri}(\mathrm{C}_\mathrm{A})$, then, the corresponding point $\mathbf{m} \in V_{\mathrm{A},\geq 0}$ satisfies $\text{supp}(\mathbf{m}) = \mathcal{I}$ and $\mathrm{A}\mathbf{m} = \boldsymbol{\xi}$. Then, Equation (18) further implies that $\log \mathbf{m} \in \mathcal{M}$, so that $\mathbf{m} = \mathbf{m}_{\boldsymbol{\xi}}$. Next assume that $\boldsymbol{\xi} \in \text{ri}(F)$, for some proper face of C_A with facial set $\mathcal{F}$. Then, by Lemma 7.3 again, the corresponding point $\mathbf{m}$ on the toric variety satisfies $\text{supp}(\mathbf{m}) = \mathcal{F}$ and the moment equations $\mathrm{A}\mathbf{m} = \boldsymbol{\xi}$. In addition, $\mathbf{m}$ is the only such point verifying $(\mathbf{w}, \boldsymbol{\mu}_\mathcal{F}) = 0$ for each $\mathbf{w} \in \text{kernel}(\mathrm{A}_\mathcal{F})$, with $\boldsymbol{\mu}_\mathcal{F} = \ln \pi_\mathcal{F}(\mathbf{m})$. Hence $\boldsymbol{\mu}_\mathcal{F} \in \pi_\mathcal{F}(\mathcal{M})$. Then $\mathbf{m}$ satisfies the conditions (14), hence conclude that $\mathbf{m} = \tau_\mathcal{F}(\exp^{\mu_F}) = \mathbf{m}_{\boldsymbol{\xi}}$, with $\tau_\mathcal{F}$ defined in (13). ■

Lemma 7.3. *For each $\boldsymbol{\xi} \in \mathrm{C}_\mathrm{A}$, $\{\mathbf{m}\} = \mathrm{P}_{\boldsymbol{\xi}} \cap V_{\mathrm{A},\geq 0}$. In fact, C_A and $V_{\mathrm{A},\geq 0}$ are homeomorphic.*

Proof. This result is derived as a direct consequence of the properties of the *moment map* (see, for example, Fulton, 1978).

The proof for the general case is given below. A different proof for the situation in which the row span of A contains the vector $\mathbf{1}_\mathcal{I}$ is given first. It identifies the MLE as the unique estimate of the natural parameter that maximizes the entropy of all distributions for which the expected value of the sufficient statistics match the observed values.

Consider the map $f : V_{\mathrm{A},\geq 0} \to \mathrm{C}_\mathrm{A}$ given by $f(\mathbf{m}) = \mathrm{A}\mathbf{m}$. It will be shown that this defines, in fact, a bijection, between $V_{\mathrm{A},\geq 0}$ and C_A. Let $\boldsymbol{\xi} \in \text{ri}(F)$ for some face (possibly improper) F of C_A and let $H : \mathbb{R}^\mathcal{I}_{\geq 0} \to \mathbb{R}$ be Shannon's entropy function $H(\mathbf{x}) = -\sum_i \mathbf{x}(i) \ln \mathbf{x}(i)$ using natural logarithm, which is continuous anywhere on its domain (at the boundary the values of H are well defined and can be obtained using continuity and the fact that $x \ln x \to 0$ for $x \downarrow 0$). The function H is strictly concave since, for any $\mathbf{x} \in \mathbb{R}^\mathcal{I}$, the hessian matrix at $\mathbf{x}$,

$$\begin{pmatrix} -\frac{1}{\mathbf{x}(i_1)} & & \\ & \ddots & \\ & & -\frac{1}{\mathbf{x}(i_r)} \end{pmatrix}$$

is negative definite, where $\text{supp}(\mathbf{x}) = \{i_j : j = 1, \ldots, r\}$.

First, it will be shown that f is surjective. From this it will follow that each non-negative vector $\mathbf{x}$ in $\mathrm{P}_{\boldsymbol{\xi}}$ will be such that $\mathbf{x}(i) = 0$ for all $i \notin \mathcal{F}$; furthermore, because of Lemma 5.6, any possible point $\mathbf{m}$ in $V_{\mathrm{A},\geq 0} \cap \mathrm{P}_{\boldsymbol{\xi}}$ satisfies $\text{supp}(\mathbf{m}) = \mathcal{F}$. The ambient space of the polytope $\mathrm{P}_{\boldsymbol{\xi}}$ is $\mathbb{R}^\mathcal{F}$ and, without loss of generality, assume that H is restricted[1] to $\mathbb{R}^\mathcal{F}$. The restriction of H to

[1] In fact, for any $\mathbf{x} \in \mathbb{R}^\mathcal{I}_F$ with $\mathbf{x}(i) = 0$ for all $i \notin \mathcal{F}$, $H(\mathbf{x}) = -\sum_{i\in\mathcal{F}} \mathbf{x}(i) \ln \mathbf{x}(i)$, which is precisely the value of H restricted to $\mathbb{R}^\mathcal{F}_{\geq 0}$.

the closed set $\mathrm{P}_{\boldsymbol{\xi}}$ (which is also compact in its ambient space) is still strictly concave and hence it achieves its unique maximum at one point $\mathbf{m}^* \in \mathrm{P}_{\boldsymbol{\xi}}$. It is not hard to see that $\mathbf{m}^* \in \mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}})$ so that $\mathrm{supp}(\mathbf{m}^*) = \mathcal{F}$. Since $\mathbf{m}^*$ is a stationary point of H and it belongs to $\mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}})$, there exists a relatively open neighborhood $B_{\mathbf{m}^*}$ of $\mathbf{m}^*$ inside $\mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}})$ such that the directional derivative of H at $\mathbf{m}^*$ is zero along each direction $\mathbf{w}$ satisfying $\mathbf{m}^* + \alpha\mathbf{w} \in B_{\mathbf{m}^*}$, $||\mathbf{w}|| = 1$, for some sufficiently small positive α. This implies that, for each $\mathbf{w} \in \mathbb{R}^{\mathcal{I}}$ such that $\pi_{\mathcal{F}}(\mathbf{w}) \in \mathrm{kernel}(\mathrm{A}_{\mathcal{F}})$,

$$0 = \sum_{i \in \mathcal{F}} (\mathbf{w}(i) \ln \mathbf{m}^*(i) + \mathbf{w}(i)) = \sum_{i \in \mathcal{F}} \mathbf{w}(i) \ln \mathbf{m}^*(i), \tag{32}$$

where $\sum_{i \in \mathcal{F}} \mathbf{w}(i) = 0$, since $\mathbf{1}_{\mathcal{F}}$ belongs to the row span of $\mathrm{A}_{\mathcal{F}}$. Equivalently, for each integer value vector $\mathbf{w} \in \mathrm{kernel}(\mathrm{A}_{\mathcal{F}})$,

$$(\mathbf{m}^*)^{\tau_{\mathcal{F}}(\mathbf{w})^+} = (\mathbf{m}^*)^{\tau_{\mathcal{F}}(\mathbf{w})^-},$$

where, as usual, $0^0 = 1$. Given the convention that $0^a = 0$ for every $a \neq 0$ and the fact that $\tau_{\mathcal{F}}(\mathrm{kernel}(\mathrm{A}_{\mathcal{F}})) \subset \mathrm{kernel}(\mathrm{A})$, the previous equation holds for all integer $\mathbf{w} \in \mathrm{kernel}(\mathrm{A})$. This implies that equation (18) is satisfied. Hence $\mathbf{m}^* \in V_{\mathrm{A},\geq 0}$.

In order to show injectivity, suppose that there exist a point $\mathbf{m}^{**} \in V_{\mathrm{A},\geq 0} \cap \mathrm{P}_{\boldsymbol{\xi}}$ distinct from $\mathbf{m}^*$. Then, $\mathrm{supp}(\mathbf{m}^{**}) = \mathcal{F}$, hence $\mathbf{m}^{**} \in \mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}})$. In fact, arguing by contradiction, if this were not the case, then necessarily, by Lemma 5.6, $\mathrm{supp}(\mathbf{m}^{**})$ would be a facial set corresponding to a face F' of F, contradicting the assumption that $\boldsymbol{\xi} \in \mathrm{ri}(F)$. Thus, $\mathbf{m}^{**} \in \mathrm{ri}(\mathrm{P}_{\boldsymbol{\xi}}) \cap V_{\mathrm{A},\geq 0}$, hence it satisfies equation (32) for all $\mathbf{w} \in \mathrm{kernel}(\mathrm{A})$. But this in turn implies that $\mathbf{m}^{**}$ is also a stationary point for the entropy function H distinct from $\mathbf{m}^*$, contradicting the strict concavity of H. Hence $\mathbf{m}^{**}$ does not exist.

If the row span of A does not contain the vector $\mathbf{1}_{\mathcal{I}}$, then the proof so far can still be used to show that there exists a homeomorphism between $V_{\geq 0,\mathrm{A}} \cap \Delta_{\mathcal{I}}$ and the polytope $\mathrm{convhull}(\mathrm{A})$, where $\Delta_{\mathcal{I}}$ is the simplex in $\mathbb{R}^{\mathcal{I}}$.

For the more general case of a design matrix A not containing in its row span the constant vectors in $\mathbb{R}^{\mathcal{I}}$, the moment map property is proved using the following result, contained in Fulton (1978).

Proposition 7.4. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be vector in $\mathbb{R}^k$ that span $\mathbb{R}^k$ and let C be the cone they span. Then, the map $F : \mathbb{R}^k \to \mathbb{R}^k$ defined by*

$$F(\mathbf{x}) = \sum_{i=1}^{n} \exp^{(\mathbf{u}_i, \mathbf{x})} \mathbf{u}_i \tag{33}$$

determines a real analytic isomorphism of $\mathbb{R}^k$ into the interior of C.

Assume for convenience, and without loss of generality, that A is of full rank and consider the following function from $\mathbb{R}^k$ into $\mathbb{R}^{\mathcal{I}}_{>0}$, obtained as a composition of the (element-wise) exponential function and the monomial map defined in equation (15): $\kappa_{\mathrm{A}}(\mathbf{x}) = \phi_{\mathrm{A}}(\exp^{\mathbf{x}}) = \exp^{\mathrm{A}^\top \mathbf{x}}$. Since the exponential function maps bijectively $\mathbb{R}^k$ into $\mathbb{R}^k_{>0}$, it is clear that $\mathrm{cl}(\mathrm{im}(\kappa_{\mathrm{A}})) = \mathrm{cl}(\mathrm{im}(\phi_{\mathrm{A}})) = V_{\mathrm{A},\geq 0}$, where the last equality is Theorem 5.5.

By Lemma 5.6, for each $\mathbf{m} \in V_{\mathrm{A},\geq 0}$, there exists a (possibly improper) face F of C_{A} such that $\mathrm{supp}(\mathbf{m}) = \mathcal{F}$. Therefore, for all such points, $\mathbf{m} = \tau_{\mathcal{F}}(\mathrm{im}(\phi_{\mathrm{A}_{\mathcal{F}}})) = \tau_{\mathcal{F}}(\mathrm{im}(\kappa_{\mathrm{A}_{\mathcal{F}}}))$, with $\mathrm{A}_{\mathcal{F}}$ being the sub-matrix of A consisting of the columns in $\mathcal{F}$. Let $\mathrm{A}^*_{\mathcal{F}}$ denote the full rank matrix with the same row span as $\mathrm{A}_{\mathcal{F}}$ and let $k_F = \mathrm{rank}(\mathrm{A}_{\mathcal{F}})$. Then, the function F in equation (33) can be expressed as

$F(\mathbf{x}) = \mathrm{A}^*_{\mathcal{F}}\mathbf{m}(\mathbf{x})$, where, for all $\mathbf{x} \in \mathbb{R}^{k_F}$, $\mathbf{m}(\mathbf{x}) = \exp^{(\mathrm{A}^*_{\mathcal{F}})^\top \mathbf{x}} \in \mathrm{im}(\kappa_{\mathrm{A}_{\mathcal{F}}})$. By Proposition 7.4, there is a one-to-one correspondence between $\mathbb{R}^{k_F}$ and $\mathrm{ri}(\mathrm{C}_{\mathrm{A}^*_{\mathcal{F}}})$ and hence between $\mathbb{R}^{k_F}$ and $\mathrm{ri}(\mathrm{C}_{\mathrm{A}_{\mathcal{F}}})$. Since, by Lemma 7.5, $\mathbb{R}^{k_F}$ is also in one-to-one correspondence with all the points $\mathbf{m} \in V_{\mathrm{A},\geq 0}$ such that $\mathrm{supp}(\mathbf{m}) = \mathcal{F}$ via the map $\mathbf{x} \to \tau_{\mathcal{F}}(\kappa_{\mathrm{A}_{\mathcal{F}}}(\mathbf{x}))$, the result follows.

As a final remark, since f is a continuous function, the moment map f defines a homeomorphism between C_{A} and $V_{\mathrm{A},\geq 0}$. ■

Lemma 7.5. *Let* A *be a full-row rank* $k \times n$ *matrix with integer entries. Then, the monomial map (15) is a homeomorphism between* $\mathbb{R}^k_{>0}$ *and* $\mathrm{im}(\phi_{\mathrm{A}})$.

Proof. The result follows immediately from the fact that A is of full row rank and the mapping $\mathbf{t} \in \mathbb{R}^k_{>0} \to \exp\{\mathrm{A}^\top \log \mathbf{t}\} = (\mathbf{t}^{\mathbf{a}_1}, \ldots, \mathbf{t}^{\mathbf{a}_n})$ is in fact a homeomorphism, being the composition of bijective continuous functions with continuous inverses. ■

8 Appendix B

Connections with some results in Haberman (1974).

- **Equivalence of Theorem 2.5 of Haberman (1974) and Theorem 3.5**
  Let A^* be a design matrix whose row span is $\mathcal{M} \ominus \mathcal{N}$ and let $\mathbf{z} = \mathrm{A}^*\mathbf{n}$. By Theorem 3.5, the MLE does not exist when $\mathbf{z}$ belongs to a face of the marginal cone generated by the columns of A^*. This occurs if and only if there exists a $\boldsymbol{\zeta}_F$ such that $(\boldsymbol{\zeta}_F, \mathbf{z}) \geq (\boldsymbol{\zeta}_F, \mathbf{z}')$ for all $\mathbf{z}' = \mathrm{A}^*\mathbf{x}$, $\mathbf{x} \in S(\mathcal{N})$, if and only if $(\boldsymbol{\zeta}_F, \mathrm{A}^*\mathbf{n}) \geq (\boldsymbol{\zeta}_F, \mathrm{A}^*\mathbf{x})$ for all $\mathbf{x} \in S(\mathcal{N})$, if and only if $(\boldsymbol{\mu}, \mathbf{n} - \mathbf{x}) \geq 0$ for all $\mathbf{x} \in S(\mathcal{N})$, where $\boldsymbol{\mu} = (\mathrm{A}^*)^\top \boldsymbol{\zeta}_F \in \mathcal{M} \ominus \mathcal{N}$.

- **Equivalence of Theorem 2.3 of Haberman (1974) and Corollary 4.7**
  Let A be a design matrix whose rows span the log-linear subspace $\mathcal{M}$. The MLE does not exist if and only if $\mathrm{A}\mathbf{n} = \mathbf{t} \in \mathrm{ri}(F)$ for some face F of the marginal cone, if and only if $\{i \in \mathcal{I} \colon \mathbf{n}(i) > 0\} \subseteq \mathcal{F}$, where $\mathcal{F}$ is the facial set associated with F, if and only if there exists a vector $\boldsymbol{\zeta}_F$ such that $\boldsymbol{\mu} = \boldsymbol{\zeta}_F^\top \mathrm{A} \in \mathcal{M}$ satisfies $\boldsymbol{\mu}_{\mathcal{F}} > \mathbf{0}$, $\boldsymbol{\mu}_{\mathcal{F}^c} = \mathbf{0}$ and $(\mathbf{n}, \boldsymbol{\mu}) = 0$.

Acknowledgments

This article is based on my doctoral dissertation work under the supervision of Stephen E. Fienberg, to whom I am grateful for his constant advice and guidance as an advisor and colleague. I am very much indebted to Isabella Verdinelli for carefully reading initial drafts of this manuscript and for many helpful comments and suggestions.

References

Agresti, A. (2002). *Categorical Data Analysis* (second ed.). New York: John Wiley & Sons.

Bardorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. New York: John Wiley & Sons.

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.

Brown, L. D. (1986). *Fundamental of Statistical Exponential Families*, Volume 9 of *IMS Lecture Notes-Monograph Series*. California: Institute of Mathematical Statistics.

Chen, Y., I. H. Dinwoodie, and S. S. (2006). Sequential importance sampling for multiway tables. *Annnals of Mathematical Statistics 34*(1).

Cover, T. M. and J. A. Thomas (1991). *Information Theory*. New York: John Wiley & Sons.

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. and F. Matúš (2001). Convex cores of measures. *Studia Scientiarum Mathematicarum Hungarica 38*, 177–190.

Csiszár, I. and F. Matúš (2003). Information projection revisited. *IEE Transaction of Information Theory 49*(6), 1474–1490.

Csiszár, I. and F. Matúš (2005). Closure of exponential families. *The Annals of Probability 33*(2), 582–600.

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.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics2 26*(1), 363–397.

Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant (2005). Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. To appear in Journal of Symbolic Computation, Special issue on Computational Algebraic Statistics.

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 (2006a). Computing Maximum Likelihood Estimates for Log-linear Models. Manuscript in preparation.

Fienberg, S. E. and A. Rinaldo (2006b). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. Accepted in Journal of Statistical Planning and Inference.

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. Technical report. To appaer in the *Annals of Statistics*. Available at http://www.research.microsoft.com.

Grunwald, P. D. and P. A. Dawid (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics 32*, 1367–1433.

Haberman (1974). *The Analysis of Frequency Data*. Chicago, Illinois: University of Chicago Press.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected counts. *The Annals of Statistics 5*(6), 1148–1169.

Jordan, M. and M. Wainwright (2003). Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley.

Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the Americal Statistical Association 81*(394), 483–493.

Lang, J. B. (1996). On the comparison of multinomial and Poisson log-linear models. *Journal of the Royal Statistical Society 1*(58), 253–266.

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.

Morris, C. (1975). Central limit theorems for multinomial sums. *The Annals of Statistics 3*(1), 165–188.

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.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton: Princeton University Press,.

Sturmfels, B. (1996). *Gröbner Bases and Convex Polytopes*. American Mathematical Society.

Sturmfels, B. (2003). Algebra & Geometry of Statistical Models. http://www-m10.mathematik.tu-muenchen.de/neumann/.

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.

Ziegler, M. G. (1998). *Lectures on Polytopes*. New York: Springer-Verlag.