

# Polyhedral Conditions for the Nonexistence of the MLE for Hierarchical Log-linear Models

Nicholas Eriksson<sup>a</sup> Stephen E. Fienberg<sup>b,c</sup> Alessandro Rinaldo<sup>b</sup>  
Seth Sullivant<sup>a</sup>

<sup>a</sup>*Department of Mathematics, University of California, Berkeley*

<sup>b</sup>*Department of Statistics, Carnegie Mellon University*

<sup>c</sup>*Center for Automated Learning and Discovery and Center for Computer and Communication Security, Carnegie Mellon University*

---

## Abstract

We provide a polyhedral description of the conditions for the existence of the maximum likelihood estimate (MLE) for a hierarchical log-linear model. The MLE exists if and only if the observed margins lie in the relative interior of the marginal cone. Using this description, we give an algorithm for determining if the MLE exists. If the tree width is bounded, the algorithm runs in polynomial time. We also perform a computational study of the case of three random variables under the no three-factor effect model.

*Key words:* maximum likelihood estimate (MLE), marginal cone, tree width, collapsing

---

## 1 Introduction

In the analysis of contingency tables using log-linear models, the maximum likelihood estimate (MLE) of the underlying parameters (or equivalently of the expectations of the cell counts) plays a fundamental role for computation, the assessment of model fit, and model interpretation. In particular, the existence of the MLE is crucial for the determination of degrees of freedom of traditional  $\chi^2$  large sample approximations (see, for example, Bishop et al.,

---

*Email addresses:* [eriksson@math.berkeley.edu](mailto:eriksson@math.berkeley.edu) (Nicholas Eriksson), [fienberg@stat.cmu.edu](mailto:fienberg@stat.cmu.edu) (Stephen E. Fienberg), [arinaldo@stat.cmu.edu](mailto:arinaldo@stat.cmu.edu) (Alessandro Rinaldo), [seths@math.berkeley.edu](mailto:seths@math.berkeley.edu) (Seth Sullivant).

1975) and for exact or approximate techniques for computing  $p$ -values. If the MLE does not exist, then the standard procedures and their approximations require alteration.

The characterizations of the conditions for the existence of the MLE developed in the statistical literature are non-constructive, in the sense that they do not directly lead to a numerical implementation (see Haberman, 1974, Appendix B). As a result, the possibility of the nonexistence of the MLE is rarely considered by practitioners and the only available indication of it is a lack of convergence of the iterative algorithms used to approximate the MLE.

The problem of nonexistence has long been known to relate to the presence of zero cell counts in the table, e.g., see Fienberg (1970); Haberman (1974); Bishop et al. (1975). Zero counts arise frequently in large sparse tables where the total sample size is small relative to the number of cells in the table, e.g., see Koehler (1976). Thus for small contingency tables with a large sample size, the nonexistence of the MLE is a relatively infrequent problem. This is because for small contingency tables (nearly) all of the cell entries in the table will be positive, which, as we will see, guarantees the existence of the MLE. However, the nonexistence of the MLE is a potentially common problem in applications in the biological, medical, and social sciences, where the contingency tables which arise are large and sparse. Unfortunately, in many such applications researchers “collapse” large sparse tables to form one of smaller dimension and/or size. As Bishop et al. (1975) and Lauritzen (1996) make clear, such collapsing can lead to erroneous statistical inferences about associations among the variables displayed in the table. Here we explore collapsing as a technical device to illustrate the combinatorial complexity that arises in studying the nonexistence of the MLE.

The goals of this paper are two-fold. First, we show that the nonexistence of the MLE is equivalent to the margins of the observed contingency table lying on a facet of the marginal cone of the underlying hierarchical log-linear model. This polyhedral reinterpretation of the problem immediately leads to easily implementable algorithms for determining whether or not the MLE exists given an observed contingency table. We discuss these algorithms in Section 3. From the practical standpoint, this characterization gives a simple way to check whether or not the MLE exists before using numerical methods to estimate the MLE. In the event the MLE does not exist, identifying those zero cell counts that cause the non-existence problem requires generalizations of the basic algorithm we describe.

The second goal of this paper is to alert the mathematical reader to a rich source of combinatorial problems that arise from statistical applications. The polyhedral cones we are concerned with have received attention in various guises (e.g., the “correlation polytope” in Deza and Laurent (1997) and the

“marginal polytope” in Jordan and Wainwright (2003)). Thus our particular problem of deciding if a point in this cone is on a facet is a new variation on an old theme. Given recent computational advances, this also suggests the problem of developing efficient algorithms for computing the convex hulls of highly symmetric polyhedra. We discuss these issues in Section 4.

The outline for this paper is as follows. In Section 2 we define hierarchical models and the MLE, and we show that the MLE exists if and only if the observed margins belong to the relative interior of a polyhedron. In Section 3 we use this fact to describe an algorithm for checking the existence of the MLE. The algorithm uses linear programming and runs in polynomial time if the tree width of the model is bounded. Section 4 focuses on the study of the complexity of the problem for 3-way tables. In particular, we consider the collapsing operation as a combinatorial tool for studying the nonexistence of the MLE.

## 2 Hierarchical models and the MLE

In this section, we introduce hierarchical models and the maximum likelihood estimate and we show that the maximum likelihood estimate exists if and only if certain polyhedral conditions are satisfied. For this and the remaining sections, we assume the reader is familiar with the basics of polyhedral geometry. Two standard references are Ziegler (1998) for basics on polyhedra and Schrijver (1998) for algorithmic aspects including linear programming. Our polyhedral condition is a reformulation of a result of Haberman (1974).

Contingency tables are collections of non-negative integers arising from cross-classifying a set of objects into categories or cells indexed by a set of labels  $d$  corresponding to variables of interest (see Bishop et al., 1975; Lauritzen, 1996). More precisely, we get a  $K$ -way contingency table  $\mathbf{n}$  by taking a sample of independent and identically distributed observations on a vector of  $K$  discrete random variables  $(X_1, \dots, X_K)$ . The  $j$ th random variable  $X_j$  takes values in the set  $[d_j] := \{1, 2, \dots, d_j\}$ . We call the various states of the random variables *levels*. Let  $d = \otimes_{j=1}^K [d_j]$ . Thus each  $i \in d$  identifies the number  $\mathbf{n}(i)$ .

Although the entries in the table  $\mathbf{n}$  are integer-valued, we treat  $\mathbf{n}$  as an element of  $\mathbb{R}^d$ , the space of all real valued functions on the multi-index set  $d$  endowed with the usual inner product  $\mathbf{x}^T \mathbf{y} = \sum_{i \in d} \mathbf{x}(i) \mathbf{y}(i)$  for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ . For the remainder of the paper, we assume that the index set  $d$  is linearized in some fashion, so that we can represent the table  $\mathbf{n}$  as a vector.

The statistical analysis of tables using log-linear models focuses on inference about parameters in a model or equivalently on inferences about the mean

vector  $\mathbf{m} = \mathbb{E}(\mathbf{n})$  of the observed table under the assumption that  $\mathbf{m} > \mathbf{0}$ , so that  $\mu = \log \mathbf{m}$  is well defined. There are interesting extensions of the ideas in this paper to situations where we know a priori that some entries of  $\mathbf{m}$  are zero (c.f., Bishop et al. (1975); Haberman (1974); Fienberg (1970)).

Log-linear models arise from assuming  $\mu \in \mathcal{M}$ , where  $\mathcal{M}$  is a  $p$ -dimensional linear subspace  $\mathcal{M} \subseteq \mathbb{R}^d$  such that  $\mathbf{1}_d \in \mathcal{M}$ . A common way of obtaining  $\mathcal{M}$  is by specifying a hierarchical model. A hierarchical model is determined by a simplicial complex  $\Delta$  on  $K$  vertices from which a 0-1 matrix  $A_\Delta$  is constructed whose rows span  $\mathcal{M}$  in the following way. Let  $\{\mathcal{F}_1, \dots, \mathcal{F}_f\}$  be the facets of  $\Delta$  and, for each  $\mathcal{F}_s$  and  $i \in d$ , let  $d_{\mathcal{F}_s} = \bigotimes_{j \in \mathcal{F}_s} [d_j]$  and  $i_{\mathcal{F}_s}$  be the restriction of  $i$  to  $d_{\mathcal{F}_s}$ . Let  $F_{\mathcal{F}_s}$  be the set of functions on  $d$  that depends on  $i$  only through  $i_{\mathcal{F}_s}$ . That is,

$$F_{\mathcal{F}_s} := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}(i) = \mathbf{x}(j) \text{ for all } i, j \text{ with } i_{\mathcal{F}_s} = j_{\mathcal{F}_s}\}.$$

Then the linear subspace  $\mathcal{M}$  corresponding to the hierarchical log-linear model  $\Delta$  takes the form

$$\mathcal{M}_\Delta = \sum_{\mathcal{F}_s \in \Delta} F_{\mathcal{F}_s}.$$

Let  $A_\Delta$  be a 0-1 matrix having dimension  $v \times |d|$ , where  $v = \sum_s \prod_{j \in \mathcal{F}_s} d_j$  and  $|d|$  is the cardinality of this index set. Each row of  $A_\Delta$  is indexed by the pair  $(\mathcal{F}_s, i_{\mathcal{F}_s})$  and is equal to the indicator function  $\chi(i_{\mathcal{F}_s})$ , a vector in  $\mathbb{R}^d$  which is 1 on coordinates  $i_{\mathcal{F}_s}$  and 0 otherwise. Then the rows of  $A_\Delta$  span  $\mathcal{M}_\Delta$ , so a hierarchical model can be identified by a collection of  $K$  levels  $\mathbf{d} = (d_1, \dots, d_K)$  and a simplicial complex  $\Delta$  on  $K$  nodes.

Data displayed in the form of contingency tables arise from various sampling schemes involving the observations on the random variables (see Bishop et al., 1975; Haberman, 1974). The results that follow are valid for the following three schemes:

**Poisson Sampling.** The total number  $n = |\mathbf{n}|$  of counts is random, where, for a non-negative vector  $\mathbf{x}$ ,  $|\mathbf{x}| = \sum_i \mathbf{x}(i)$ , and the counts are in fact independent Poisson random variables.

**Multinomial sampling.** The total number  $n = |\mathbf{n}|$  of counts is fixed by design.

**Product Multinomial sampling.** Let  $\mathcal{B} \subset \{1, \dots, n\}$  and  $d_{\mathcal{B}} = \bigotimes_{j \in \mathcal{B}} d_j$ , as above. For each  $b \in \mathcal{B}$ , the number of counts  $|\mathbf{n}(i_b)|$  is fixed by design. Here, we assume, as is commonly done in the statistical literature, that  $\mathcal{B}$  is always a face of  $\Delta$ .

Given a table  $\mathbf{n}$  on the fixed set of levels  $\mathbf{d} = (d_1, \dots, d_K)$  and a simplicial complex  $\Delta$ , the maximum likelihood estimate of  $\mu$  is the point  $\hat{\mu} \in \mathcal{M}_\Delta$  such that  $\hat{\mathbf{m}} \equiv \exp(\hat{\mu})$  best approximates the unknown mean  $\mathbf{m} = \mathbb{E}(\mathbf{n})$  in the sense that it maximizes the probability of observing the actual table  $\mathbf{n}$ , i.e.,

joint distribution of the counts  $\mathbf{n}$  as a function of the mean vector  $\mathbf{m}$ . This probability is also known as the likelihood function when we express it as a function of the parameters  $\mathbf{m}$  given the data  $\mathbf{n}$ . The log-likelihood function  $\ell(\mathbf{m})$  is the logarithm of the likelihood function.

For a given observed table  $\mathbf{n}$ , we can write the log-likelihood as:

$$\ell(\mathbf{m}) = \log \Pr(\mathbf{n}(i) \mid \mathbf{m}(i), i \in d) = \sum_{i \in d} \mathbf{n}(i) \log \mathbf{m}(i) - \sum_{i \in d} \mathbf{m}(i) + C_{\mathbf{n}}$$

where  $C_{\mathbf{n}}$  is the logarithm of the normalization constant and depends only on  $\mathbf{n}$  and the particular sampling scheme. For a hierarchical model  $\Delta$ , we can reparametrize the log-likelihood as:

$$\ell(\mu) = (\mathcal{P}_{\Delta} \mathbf{n})^T \mu - \sum_{i \in d} \exp(\mu(i)) + C_{\mathbf{n}}$$

where  $\mathcal{P}_{\Delta}$  is the projection matrix onto  $\mathcal{M}_{\Delta}$ .

The *maximum likelihood estimate*, or MLE, of  $\mu$  is the vector  $\hat{\mu} \in \mathcal{M}_{\Delta}$  such that:

$$\ell(\hat{\mu}) = \sup_{\mu \in \mathcal{M}_{\Delta}} \ell(\mu)$$

**Definition 1** *We say that the MLE does not exist or is undefined if the supremum is not attained by a finite vector  $\hat{\mu}$ .*

The log-likelihood depends on the observed table  $\mathbf{n}$  only through  $\mathcal{P}_{\Delta} \mathbf{n}$  or, equivalently, since the rows of  $A_{\Delta}$  span  $\mathcal{M}_{\Delta}$ , the vector  $\mathbf{t} = A_{\Delta} \mathbf{n}$ . Therefore, in order to establish the existence and find the numerical value of the MLE, we need only observe  $\mathbf{t}$ , the vector of margins of the observed table; these are known as the *minimal sufficient statistics* for the model.

The requirement that the MLE must be finite derives from the assumption of positivity of the coordinates of the mean vector  $\mathbf{m}$  and the issue of existence translates into the problem of existence of a *strictly positive* vector  $\mathbf{m}$  maximizing  $\ell(\mathbf{m})$ . This means that there are no cells that are known to have zero probability a priori, i. e. no structural zeroes. Haberman (1974) and Lauritzen (1996), considered extended classes of log-linear models in which the MLE is the limit of points in  $\mathcal{M}_{\Delta}$  realizing the supremum of the log-likelihood function. This would require some of the coordinates of the estimated mean vector to be zero. The present work is concerned only with the conditions guaranteeing the existence of a strictly positive solution.

Surprisingly, the study of the conditions of existence of the MLE has received only limited attention in the statistical literature. Essentially all available results are variations of the following theorem due to Haberman (1974):

**Theorem 2** *Under any of the three sampling schemes described above, a necessary and sufficient condition for the existence of the MLE is that there exists  $\mathbf{z} \in \ker(A_\Delta)$  such that  $\mathbf{n} + \mathbf{z} > \mathbf{0}$ .*

For a strengthening of Theorem 2 see Geiger et al. (2002). For a given log-linear model  $\Delta$ , define the marginal cone  $P_\Delta = C(A_\Delta)$  to be the set of minimal sufficient margins,  $\mathbf{t}$ , where, for any matrix  $A$ ,  $C(A)$  indicates the cone generated by its columns. Let  $\text{relint}(P_\Delta)$  denote the relative interior of  $P_\Delta$ , defined as the interior of  $P_\Delta$  with respect to its embedding into the smallest linear space containing it. Then, the following corollary provides a polyhedral reinterpretation of the conditions for the existence of the MLE:

**Corollary 3** *Under any of the three sampling schemes, the MLE for the mean vector  $\mathbf{m}$  exists if and only if the margins  $\mathbf{t} = A_\Delta \mathbf{n}$  belong to  $\text{relint}(P_\Delta)$ .*

**PROOF.** A vector of margins  $\mathbf{t}$  lies in the relative interior of the polyhedral cone  $P_\Delta$  if and only if there is a table  $\mathbf{x}$  with strictly positive cells such that  $A\mathbf{x} = \mathbf{t}$ . Theorem 2 then implies that the MLE exists if and only if  $\mathbf{t} \in \text{relint}(P_\Delta)$ .  $\square$

### 3 Determining the existence of the MLE

In this section, we describe algorithms for determining whether the MLE for a given table  $\mathbf{n}$  and model  $\Delta$  exists. To make the mathematical statements in this section concise, we assume that  $A_\Delta$  contains extra rows determined by the faces of  $\Delta$  in addition to those rows determined by the facets of  $\Delta$ . Since this over-parameterization does not change the row span  $\mathcal{M}_\Delta$ , the matrix  $A_\Delta$  describes the same hierarchical log-linear model. To implement the algorithms we describe, one can relax this condition on  $A_\Delta$ .

By Corollary 3, the maximum likelihood estimate does not exist if and only if the vector of observed margins  $\mathbf{t} = A_\Delta \mathbf{n}$  lies on a facet of  $P_\Delta$ . Hence, we want to show that there is a nontrivial vector  $\mathbf{c}$  in the dual cone of  $P_\Delta$  which attains its maximum value at  $\mathbf{t}$  but does not attain its maximum value at some other point of  $P_\Delta$ . The existence of such a  $\mathbf{c}$  implies that  $\mathbf{t}$  lies on a facet of  $P_\Delta$ . However, this can be decided by determining if the polyhedral cone

$$F_{\mathbf{n}}^\Delta = \{\mathbf{c} \mid \mathbf{c}^T A_\Delta \leq \mathbf{1}^T \cdot \mathbf{c}^T \mathbf{t}\} \quad (1)$$

contains only those vectors orthogonal to the linear hull of  $P_\Delta$ .

Note that this linear system involves exponentially many inequalities in the number of random variables  $K$ . We show, however, that if the model  $\Delta$  satisfies

certain nice complexity properties, the linear system (1) has an equivalent formulation using only polynomially many inequalities. Since we can solve linear programs in polynomial time (e.g., Schrijver, 1998), this implies the following result:

**Theorem 4** *There is an algorithm for deciding the triviality of the linear program (1) which runs in polynomial time in the size of the input data and the number of levels of each random variable whenever the simplicial complex  $\Delta$  has bounded tree width.*

A precise formulation appears in Theorem 9 where the exact complexity bounds are stated. First, we define all the objects in question.

**Definition 5** *A simplicial complex  $\Delta$  is reducible if there is a decomposition of  $\Delta$  into  $(\Delta_1, S, \Delta_2)$  such that*

- (1)  $\Delta_1 \cup \Delta_2 = \Delta$ ,
- (2)  $|\Delta_1| \cap |\Delta_2| = S$ , and
- (3)  $S \in \Delta_1$  and  $S \in \Delta_2$ .

Here  $|\Delta_i|$  denotes the underlying set of  $\Delta_i$ . A simplicial complex is called decomposable or chordal if it is reducible and each of  $\Delta_1$  and  $\Delta_2$  are either decomposable or a simplex.

**Definition 6** *The tree width  $T(\Delta)$  of a simplicial complex  $\Delta$  is one less than the size of the maximal face in the smallest decomposable complex that contains  $\Delta$ . That is,*

$$T(\Delta) = \min_{\Delta \subset \Gamma} \max_{C \in \Gamma} |C| - 1$$

where the minimum runs over all decomposable  $\Gamma$  with all faces of  $\Delta$  in  $\Gamma$ . A decomposable simplicial complex  $\Gamma$  that attains the minimum is called a chordal triangulation of  $\Delta$ .

For instance, the tree width of the  $K$ -cycle,  $\Delta = [12][23] \cdots [(K-1)K][1K]$ , is always 2 since a  $K$ -cycle does not have tree width 1 (i.e., it is not a tree), and the simplicial complex  $\Gamma = [123][134] \cdots [1(K-1)K]$  is a decomposable complex that triangulates the  $K$ -cycle. We study the  $K$ -cycle in more detail in Example 10 below. Note that the tree width of a simplicial complex  $\Delta$  only depends on the structure of its underlying graph because every decomposable simplicial complex is determined by its 1-skeleton. Graphs with bounded tree width are natural families to consider when one wishes to bound the complexity of algorithms related to the underlying graphs. See for example (Jordan and Wainwright, 2003) for applications to directed and undirected graphical models.

The proof of Theorem 4 follows from a series of results relating the system of linear inequalities to systems of inequalities for chordal triangulations. Our

goal is to produce a polyhedral cone whose triviality is equivalent to the triviality of the cone (1) but whose description involves fewer linear equations and inequalities.

**Lemma 7** *Suppose that  $\Gamma$  is a model with  $\Delta \subseteq \Gamma$ . Then*

$$F_{\mathbf{n}}^{\Delta} = \pi(F_{\mathbf{n}}^{\Gamma} \cap \{\mathbf{c} \mid \mathbf{c}^F = 0 \text{ with } F \in \Gamma \setminus \Delta\}),$$

where  $\pi$  is the coordinate projection of  $F_{\mathbf{n}}^{\Gamma}$  to the ambient space of  $F_{\mathbf{n}}^{\Delta}$ . The notation  $\mathbf{c}^F$  denotes the part of the vector  $\mathbf{c}$  which is naturally labeled by the face  $F \in \Gamma$ .

**PROOF.** By definition.  $\square$

Suppose that  $\Delta$  is reducible, with decomposition  $(\Delta_1, S, \Delta_2)$ . From the vector  $\mathbf{n}$  we can compute the margins with respect to  $|\Delta_1|$  and  $|\Delta_2|$ , which we denote by  $\mathbf{n}_1$  and  $\mathbf{n}_2$ .

**Lemma 8** *Suppose that  $\Delta$  is reducible, with decomposition  $(\Delta_1, S, \Delta_2)$  and let  $\mathbf{n}$  be a table. Then*

$$F_{\mathbf{n}}^{\Delta} = \iota_1(F_{\mathbf{n}_1}^{\Delta_1}) + \iota_2(F_{\mathbf{n}_2}^{\Delta_2})$$

where the “+” indicates the Minkowski addition of the two cones and  $\iota_1, \iota_2$  are the natural embeddings of  $F_{\mathbf{n}_1}^{\Delta_1}$  and  $F_{\mathbf{n}_2}^{\Delta_2}$  into the ambient space of  $F_{\mathbf{n}}^{\Delta}$ .

**PROOF.** Modulo the lineality space of  $F_{\mathbf{n}}^{\Delta}$ , the extreme rays of  $F_{\mathbf{n}}^{\Delta}$  are precisely the facet defining inequalities of  $P_{\Delta}$  on which  $\mathbf{t}$  lies. To show the claim, it suffices to show that every facet of  $P_{\Delta}$  comes from a facet of  $P_{\Delta_1}$  or  $P_{\Delta_2}$ , in the sense that  $\text{dual}(P_{\Delta}) = \iota_1(\text{dual}(P_{\Delta_1})) + \iota_2(\text{dual}(P_{\Delta_2}))$ . But this amounts to showing that we can decide the consistency of margins for a reducible model by checking consistency for both component models,  $\Delta_1$  and  $\Delta_2$ . Now if the margins  $\mathbf{t}_1$  and  $\mathbf{t}_2$  are consistent with respect to  $\Delta_1$  and  $\Delta_2$  respectively, there are tables  $\mathbf{n}_1$  and  $\mathbf{n}_2$  such that  $A_{\Delta_1}\mathbf{n}_1 = \mathbf{t}_1$  and  $A_{\Delta_2}\mathbf{n}_2 = \mathbf{t}_2$ . Then  $\mathbf{n}_1$  and  $\mathbf{n}_2$  are margins of the decomposable model  $\Delta^* = [|\Delta_1|][|\Delta_2|]$  which satisfy the linear consistency relation that their  $S = |\Delta_1| \cap |\Delta_2|$  margins agree. Thus,  $\mathbf{t}$  are consistent  $\Delta$  marginals by Lauritzen (1996). This completes the proof.  $\square$

The description of  $F_{\mathbf{n}}^{\Delta}$  as a Minkowski sum in Lemma 8 does not give a description of  $F_{\mathbf{n}}^{\Delta}$  that is short in terms of having few facets. The key to such a short description is to recall that the Minkowski sum of two polyhedra  $P+Q$ , is the image of  $P \times Q$  under the map  $\pi$  that sends  $(x, y)$  to  $x+y$ . In particular, various properties of  $P+Q$  can be determined by studying properties of  $P \times Q$ . If  $P$  has  $m$  facets and  $Q$  has  $n$  facets, then  $P \times Q$  has only  $m+n$  facets. This

implies that if  $P$  and  $Q$  have short descriptions in terms of few facets, then so does  $P \times Q$ . Lastly, linear conditions on  $P + Q$  lift to linear conditions on  $P \times Q$ . Thus we can decide if  $(P + Q) \cap L$  is empty by considering  $(P \times Q) \cap L'$  where  $L' = \pi^{-1}(L)$ . If we accumulate all of these ideas, together with the preceding lemmas, we get the following explicit version of Theorem 4.

**Theorem 9** *Let  $\Delta$  be a simplicial complex and  $\Gamma$  a chordal triangulation of  $\Delta$ , with facets  $\Gamma_1, \dots, \Gamma_s$ . Denote by  $\mathbf{n}_t$  the  $\Gamma_t$  margin of  $\mathbf{n}$ . Then the polyhedron  $F_{\mathbf{n}}^\Delta$  is equal to the orthogonal complement of the linear hull of  $P_\Delta$  if and only if the polyhedron*

$$(F_{\mathbf{n}_1}^{\Gamma_1} \times \dots \times F_{\mathbf{n}_s}^{\Gamma_s}) \cap \{(\mathbf{c}_1, \dots, \mathbf{c}_s) \mid \sum_{i=1}^d \mathbf{c}_i^F = \mathbf{0} \text{ for all } F \in \Gamma \setminus \Delta\} \quad (2)$$

is a linear space. Furthermore, the linear description of (2) requires  $O(D^{T(\Delta)+1}s)$  equations and inequalities in an ambient space of dimension  $O(D^{T(\Delta)+1}s)$  where  $D = \max_i d_i$ . If the tree width of  $\Delta$  is bounded, the complexity of the resulting linear program is polynomial in  $K$  and  $d_i$  and the bit complexity the  $\mathbf{n}_t$ .

**PROOF.** This is straightforward once we unravel all of the definitions. The main point is that (2) projects, under the ‘‘Minkowski summation’’ map, onto  $F_{\mathbf{n}}^\Delta$ . This is because the set on the left of the  $\cap$  projects onto  $F_{\mathbf{n}}^\Gamma$  and the set on the right of the  $\cap$  is the pullback of the linear conditions which are forced in Lemma 7.

Now we will prove the statement about the complexity of the resulting linear program. Each of the sets  $F_{\mathbf{n}_t}^{\Gamma_t}$  has a description in terms of at most  $O(D^{|\Gamma_t|})$  equations and inequalities in  $O(D^{|\Gamma_t|})$  variables since it is the normal cone to a point on a simplex in an ambient space of dimension  $O(D^{|\Gamma_t|})$ . Taking the conjunction of all these equations and inequalities yields  $O(D^{T(\Delta)+1}s)$  equations and inequalities in  $O(D^{T(\Delta)+1}s)$  variables which describe the set on the left of the  $\cap$  in (2). Now, each facet  $F \in \Gamma \setminus \Delta$  contributes  $O(D^{|F|})$  equations. For each  $r$  there are  $O(\binom{T(\Delta)+1}{r}s)$  facets  $F \in \Gamma \setminus \Delta$  with  $|F| = r$  so this yields a total of

$$\sum_{r=1}^{T(\Delta)+1} O\left(\binom{T(\Delta)+1}{r} D^r s\right) = O((D+1)^{T(\Delta)+1} s) = O(D^{T(\Delta)+1} s)$$

equations on the righthand side of the  $\cap$  in (2). Thus the total number of equations and inequalities needed to describe (2) is also  $O(D^{T(\Delta)+1}s)$ . If the tree width of  $\Delta$  is bounded this expression is polynomial in  $D$  and  $K$  since  $s \leq K$  for a decomposable complex  $\Gamma$ .  $\square$

**Example 10 (5-cycle)** Now we will describe our construction in the special case where  $K = 5$  and  $\Delta$  is the 5-cycle. Let  $\Delta = [12][23][34][45][15]$  and let  $\Gamma = [123][134][145]$  be a chordal triangulation. Clearly,  $\Delta$  has tree width 2 as we previously stated. Now we construct the system of inequalities and equations in Theorem 9 for  $\Delta$  with respect to  $\Gamma$ .

The three facets of  $\Gamma$  are  $\Gamma_1 = [123]$ ,  $\Gamma_2 = [134]$ , and  $\Gamma_3 = [145]$ . From the data, we compute the matrices  $A_{\Gamma_t}$ . we determine each of the cones  $F_{\mathbf{n}_t}^{\Gamma_t}$  by the polynomially many inequalities given by

$$F_{\mathbf{n}_t}^{\Gamma_t} = \{\mathbf{c}_t \mid \mathbf{c}_t^T A_{\Gamma_t} \leq \mathbf{1}^T \cdot \mathbf{c}_t^T A_{\Gamma_t} \mathbf{n}_t\}. \quad (3)$$

For each  $t$ , the vector  $\mathbf{c}_t$  divides into blocks, one for each face  $F$  of  $\Gamma_t$ . Thus, when  $\Gamma_{t_1}$  and  $\Gamma_{t_2}$  have a nontrivial overlap, there will be some blocks,  $\mathbf{c}_{t_1}$  and  $\mathbf{c}_{t_2}$ , labeled by the same faces. For instance,  $\Gamma_1$  and  $\Gamma_2$  intersect in the face  $[13]$ .

The conjunction of all the inequalities in (3) gives all the inequalities from the description in (2). To deduce the equations, we must set to zero all of the  $\mathbf{c}_t$  block corresponding faces of  $\Gamma$  that are not in  $\Delta$  after the projection. This amounts to adding the five sets of equations:

$$\begin{aligned} \mathbf{c}_1^{[123]} = \mathbf{0}, \mathbf{c}_2^{[134]} = \mathbf{0}, \mathbf{c}_3^{[145]} = \mathbf{0}, \\ \mathbf{c}_1^{[13]} + \mathbf{c}_2^{[13]} = \mathbf{0}, \text{ and } \mathbf{c}_2^{[14]} + \mathbf{c}_3^{[14]} = \mathbf{0}. \end{aligned}$$

Alltold, we have a system of  $O(D^3)$  inequalities and equations in  $O(D^3)$  decision variables, where  $D = \max\{d_1, \dots, d_5\}$ , to decide if the cone is a linear space (as opposed to  $O(D^5)$  in  $O(D^2)$  variables in the standard representation). For a general  $K$ -cycle, Theorem 9 produces a system of  $O(D^3 K)$  inequalities and equations in  $O(D^3 K)$  variables, instead of  $O(D^K)$  inequalities and equations in  $O(D^2 K)$  variables.

## 4 Three-way tables

### 4.1 Collapsing

In this section, we let  $\Delta$  be the simplicial complex  $[12][13][23]$  on three random variables with levels  $p, q, r$ , corresponding to the log-linear model of no three-factor effect (also referred to as no second-order interaction). This is the hierarchical log-linear model on the fewest number of random variables where the facet structure of the marginal cone is not completely understood. From

a practical standpoint, the linear programming based algorithm from Section 3 runs in polynomial time to determine whether or not the MLE exists for a given table under the no three-factor effect model. However, having an understanding of the facet structure of the marginal cone provides insight into the different possible ways that the MLE might not exist. Even in this small hierarchical model, the marginal cone is quite complicated.

Denote by  $P_{\Delta}^{p,q,r} = P_{\Delta}$  the marginal cone for this model. We now place special emphasis on the levels and we seek to understand the combinatorial structure of the set of facets of  $P_{\Delta}^{p,q,r}$ . Our main tool is collapsing the  $p \times q \times r$  table to a table with fewer levels through the combination of levels.

An *elementary* collapsing of  $P_{\Delta}^{\mathbf{d}}$  is a linear transformation  $\pi: P_{\Delta}^{\mathbf{d}} \rightarrow P_{\Delta}^{\mathbf{d}'}$  which is obtained by replacing some random variable  $X_j$  and a set  $S$  of states of  $X_j$  by a new random variable  $X'_j$  with  $d_j - |S| + 1$  states where all the states in  $S$  are mapped to a single state. A collapsing is any linear map  $\pi: P_{\Delta}^{\mathbf{d}} \rightarrow P_{\Delta}^{\mathbf{d}'}$  obtained by a sequence of elementary collapsings. Collapsing occurs naturally in applications where one wishes to make coarser distinctions on the states of random variables. For instance, a random variable which represents the height of individuals might be collapsed to the binary random variable whose two states are “tall” and “short”.

Since a collapsing  $\pi$  maps  $P_{\Delta}^{\mathbf{d}}$  onto  $P_{\Delta}^{\mathbf{d}'}$ , for any facet  $F'$  of  $P_{\Delta}^{\mathbf{d}'}$ ,  $F = \pi^{-1}(F')$  is a face of  $P_{\Delta}^{\mathbf{d}}$ . If  $F$  is a facet of  $P_{\Delta}^{\mathbf{d}}$ , we say that  $F$  is obtained by collapsing the  $d_1 \times \cdots \times d_n$  table to a  $d'_1 \times \cdots \times d'_n$  table. As an example of this construction, we use collapsing to derive exponential lower bounds on the number of facets of the marginal cone of the no three-factor effect model.

**Proposition 11** *The number of facets of  $P_{\Delta}^{p,q,r}$  is at least*

$$\frac{1}{2}(2^p - 2)(2^q - 2)(2^r - 2) + pq + qr + pr.$$

**PROOF.** Up to symmetry, the facets of a  $2 \times 2 \times 2$  table are given by the conditions:

$$\begin{array}{c|c} 0 & * \\ \hline 0 & * \end{array} \quad \text{or} \quad \begin{array}{c|c} 0 & * \\ \hline * & * \end{array}$$

The  $0/*$  notation means that the facet is given by the conditions that the “0” entries in the table are zero and the  $*$  entries are non-negative. That is, the facet described by a  $0/*$  pattern is the cone over the extreme rays of the marginal cone which are marked with a  $*$ .

The first condition says that one entry in one of the margins is zero. There are  $pq + qr + pr$  margins for a  $p \times q \times r$  table. For the second condition, any  $p \times q \times r$  table can be collapsed to a  $2 \times 2 \times 2$  table in  $(2^{p-1} - 1)(2^{q-1} - 1)(2^{r-1} - 1)$  ways. Each of these collapsings gives a distinct face of  $P_{\Delta}^{p,q,r}$  of the second type

in 4 different ways. We now show that this face is in fact a facet. For this it suffices that the dimension of the linear span of the extreme rays of  $P_{\Delta}^{p,q,r}$  that are contained in this face has dimension one less than the dimension of the marginal cone. This in turn will be implied by showing that the linear span of these extreme rays together with any other extreme ray not in the face contains the entire marginal cone  $P_{\Delta}^{p,q,r}$ . Without loss of generality, by applying the natural symmetry of this problem, it follows that the extreme rays not contained in the face  $F$  are those that have indices (i.e., positions in the  $p \times q \times r$  array) in the set

$$I = \{(i_1, i_2, i_3) \mid i_1 \leq k_1, i_2 \leq k_2, i_3 \leq k_3\} \cup \{(i_1, i_2, i_3) \mid i_1 > k_1, i_2 > k_2, i_3 > k_3\},$$

for some fixed values  $k_1, k_2$ , and  $k_3$ . We denote the extreme ray indexed by  $(i_1, i_2, i_3)$  by  $e_{i_1 i_2 i_3}$ . Without loss of generality, we may take  $e_{111}$  to be the extreme ray not contained in  $F$ , by again applying the symmetry of the cone. Then for any index  $(j_1, j_2, j_3)$  with  $j_i > k_i$  for  $i = 1, 2, 3$ , we have the relation

$$e_{111} + e_{1j_2j_3} + e_{j_11j_3} + e_{j_1j_21} - e_{11j_3} - e_{1j_21} - e_{j_111} = e_{j_1j_2j_3}.$$

Since all the extreme rays on the left hand side are contained in  $F \cup \{e_{111}\}$ , this implies that  $e_{j_1j_2j_3}$  is contained in the linear span of  $F \cup \{e_{111}\}$ . By symmetry, all the extreme rays indexed by elements of  $I$  are contained in the linear span of  $F \cup \{e_{111}\}$ . This completes the proof that  $F$  is a facet.  $\square$

The cones  $P_{\Delta}^{p,q,r}$  appear in other guises in the mathematical literature. For example, Vlach (1986) studied conditions for the non-emptiness of the three-dimensional transportation polytopes. A three-dimensional transportation polytope is a set of tables

$$P_{\mathbf{t}} = \{\mathbf{x} \in \mathbb{R}_{\geq 0}^d \mid A_{\Delta}^{p,q,r} \mathbf{x} = \mathbf{t}\},$$

which is nonempty if and only if  $\mathbf{t} \in P_{\Delta}^{p,q,r}$ . Hence, his results can be reinterpreted in our language. One such result is:

**Proposition 12** *All facets of  $P_{\Delta}^{2,q,r}$  are obtained by collapsing to  $P_{\Delta}^{2,2,2}$ .*

Notice that Propositions 11 and 12 combine to show that there are exactly  $(2^q - 2)(2^r - 2) + 2(q + r) + qr$  facets of  $P_{\Delta}^{2,q,r}$ .

## 4.2 Computations

The polyhedron  $P_{\Delta}^{p,q,r}$  is given by the positive hull of the columns of  $A_{\Delta}$  as a cone with  $pqr$  extreme rays in  $\mathbb{R}^{pq+pr+qr}$ . Some of the rows of  $A_{\Delta}$  are redundant: the cone is  $pq + pr + qr - p - q - r + 1$  dimensional. It is generally a difficult

computational problem to take convex/positive hulls in a high dimensional space. The best algorithms for computing the convex hull of  $n$  points in  $\mathbb{R}^d$  take  $O(n^{\lfloor d/2 \rfloor})$  time. Using the software `polymake` by Gawrilow and Joswig (2000) we have computed the facets for a number of examples.

The group  $S_p \times S_q \times S_r$  provides a natural action on the set of facets of  $P_{p,q,r}$  given by permuting the levels of each random variable. After computing all the facets, we computed orbits under this action, which gives a better picture of the set of facets. The results of our computations are displayed in Table 1.

It is an interesting computational problem to use this very large symmetry group to better compute the convex hull. The set of symmetry classes of facets is small, and many of these classes come from collapsing from a smaller table. Thus many of the facets are known “for free” and this information should be used to compute the other facets. Also, the symmetry group is transitive on the extreme rays of the cone, so in principle one could hope to compute all the facets incident to a single extreme ray, and then use symmetry to recover the entire cone.

Given Proposition 12, a natural conjecture is that all facets are obtained by collapsing to binary tables. Our computations show that the situation is remarkably more complicated, and not all facets of  $P_{\Delta}^{p,q,r}$  for general  $p, q, r$  are obtained by collapsing.

**Example 13 (A non-collapsible facet)** *The following is a facet of  $P_{\Delta}^{4,4,4}$  that does not arise from collapsing to any smaller table.*

$$\begin{array}{cccc|cccc|cccc|cccc}
 0 & 0 & 0 & * & * & 0 & 0 & * & * & * & 0 & * & * & * & * \\
 0 & 0 & * & * & * & 0 & * & * & * & * & * & * & 0 & 0 & * & 0 \\
 0 & * & * & * & * & * & * & * & 0 & * & 0 & 0 & 0 & * & * & 0 \\
 * & * & * & * & * & 0 & 0 & 0 & * & * & 0 & 0 & * & * & * & 0
 \end{array}$$

*This example was found after examining the 39 symmetry classes of facets of  $P_{\Delta}^{4,4,4}$ .*

Based on our computations (see Table 1), we are led to the following conjecture.

**Conjecture 14** *Suppose that  $p \leq q \leq r$ . Then all facets of  $P_{\Delta}^{p,q,r}$  are obtained by collapsing from facets of  $P_{\Delta}^{p,q,q}$ .*

In general, it is true that if we fix  $p$  and  $q$ , there exists an  $r$  such that for all  $r' \geq r$ , all facets of  $P_{\Delta}^{p,q,r'}$  are obtained by collapsing from facets of  $P_{\Delta}^{p,q,r}$ . This

Table 1

Summary of Computations. The column “Orbits” counts the number of  $S_p \times S_q \times S_r$  orbits of facet types. The column “Collapsing” shows the smallest table such that all facets of  $P_{\Delta}^{p,q,r}$  are obtained by collapsing to it.

p	q	r	Dim	Extreme rays	Facets	Orbits	Collapsing
2	2	2	7	8	16	4	2 2 2
2	2	3	10	12	28	4	2 2 2
2	2	4	13	16	48	5	2 2 2
2	3	3	14	18	57	5	2 2 2
2	3	4	18	24	110	6	2 2 2
3	3	3	19	27	207	8	3 3 3
3	3	4	24	36	717	10	3 3 3
3	3	5	29	45	2379	13	3 3 3
3	3	6	34	54	7641	17	3 3 3
3	3	7	39	63	23991	20	3 3 3
3	4	4	30	48	4948	16	3 4 4
3	4	5	36	60	29387	24	3 4 4
3	4	6	42	72	153858	35	3 4 4
3	5	5	43	75	306955	42	3 5 5
4	4	4	37	64	113740	39	4 4 4

follows by noting that in a facet not obtained by collapsing, no two slices can have the same 0/\* pattern. Since for fixed  $p$  and  $q$  there are only finitely many patterns, the statement follows. Conjecture 14 merely asserts that the minimal such  $r$  is  $q$ . A natural related question is to ask how this finite complexity property of the facial structure relates to the finite complexity properties of Markov bases proved in (Santos and Sturmfels, 2003).

## 5 Summary

We have given a polyhedral description of the statistical problem of determining the existence or nonexistence of the maximum likelihood estimate for a hierarchical log-linear model for a multi-way contingency table. The computational implementation of this description in principle allows statisticians to explore for the first time the implication of patterns of zeros in large sparse tables that lead to nonexistence and thus to recast the estimation problem in terms of extended log-linear models for a corresponding incomplete con-

tingency table (c.f. Haberman, 1974)). There are further ties to this extended estimation problem inherent in the algebraic geometry description of log-linear models in terms of Gröbner bases given by Geiger et al. (2002).

## Acknowledgments

Nicholas Eriksson was supported by an NDSEG fellowship. Stephen Fienberg and Alessandro Rinaldo were supported in part by National Science Foundation Grant No. EIA-0131884 to the National Institute of Statistical Sciences and Stephen Fienberg was also supported by the Centre de Recherche en Economie et Statistique of the Institut National de la Statistique et des Études Économiques, Paris, France.

## References

- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts.
- Deza, M. M. and Laurent, M. (1997). *Geometry of Cuts and Metrics*. Springer Verlag, Berlin.
- Fienberg, S. E. (1970). Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *J. Amer. Statist. Assoc.* 65 (332), 1610–1616.
- Geiger, D., Meek, C., and Sturmfels B. (2002). On the toric algebra of graphical models, Microsoft Research. Manuscript available at <http://www.research.microsoft.com>.
- Haberman, S. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.
- Jordan, M. and Wainwright, M. (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. *J. Amer. Statist. Assoc.* 81 (394), 483–493.
- Lauritzen, S. F. (1996). *Graphical Models*. Oxford University Press, New York.
- Gawrilow, E. and Joswig, M. (2000). Polymake: a framework for analyzing convex polytopes. In: Kalai, G., Ziegler, G. M. (Eds.), *Polytopes — Combinatorics and Computation*. Birkhäuser, pp. 43–74.
- Santos, F. and Sturmfels, B. (2003). Higher Lawrence configurations. *Journal of Combinatorial Theory, Series A*, 103, 151–164.
- Schrijver, A. (1998). *Theory of Integer and Linear Programming*. John Wiley & Sons, New York.

- Vlach, M. (1986). Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Disc. Appl. Math.* 13, 61-78.
- Ziegler, G. (1998). *Lectures on Polytopes*. GTM 152, Springer-Verlag, New York.