Fitting log-linear models in sparse contingency tables using the eMLEloglin R package

Matthew Friedlander

Abstract

Log-linear modeling is a popular method for the analysis of contingency table data. When the table is sparse, and the data falls on a proper face F of the convex support, there are consequences on model inference and model selection. Knowledge of the cells determining F is crucial to mitigating these effects. We introduce the R package (R Core Team (2016)) eMLEloglin for determining F and passing that information on to the glm() function to fit the model properly.

1 Introduction

Data in the form of a contingency table arise when individuals are cross classified according to a finite number of criteria. Log-linear modeling (see e.g., Agresti (1990), Christensen (1997), or Bishop et al. (1975)) is a popular and effective methodology for analyzing such data enabling the practitioner to make inferences about dependencies between the various criteria. For hierarchical log-linear models, the interactions between the criteria can be represented in the form of a graph; the vertices represent the criteria and the presence or absence of an edge between two criteria indicates whether or not the two are conditionally independent (Lauritzen (1996)). This kind of graphical summary greatly facilitates the interpretation of a given model.

Log-linear models are typically fit by maximum likelihood estimation (i.e. we attempt compute the MLE of the expected cell counts and log-linear parameters). It has been known for many years that problems arise when the sufficient statistic falls on the boundary of the convex support, say C, of the model (Feinberg and Rinaldo (2007)). This generally occurs in sparse contingency with many zero cells. In such cases, algorithms for computing the MLE can fail to converge. Moreover, the effective model dimension will be reduced and the degrees of freedom of the usual goodness of fit statistics will be incorrect. Only fairly recently, have algorithms been developed to begin to deal with this situation (see Eriksson et al. (2006), Geyer (2009) and Feinberg and Rinaldo (2012)). It turns out that identification of the face F of C containing the data in its relative interior is crucial to efficient and reliable computation of the MLE and of the effective model dimension. If $F = C$ then the MLE exists and it's calculation is straightforward. If not (i.e. $F \subset C$), the log-likelihood has its maximum on the boundary and remedial steps must be taken to find and compute those parameters that can be estimated.

The outline of this paper is as follows. In Section 2, we describe necessary and sufficient conditions for the existence of the MLE. In Section 3, we place these conditions in the context of convex geometry. In Section 4, we describe a linear programming algorithm to find F. We then discuss how to compute ML estimates and find the effective model dimension. In Section 5, we introduce the eMLEloglin R package for carrying out the tasks described in Section 4.

2 Conditions for the existence of the MLE

Let V be a finite set of indices representing $|V|$ criteria. We assume that the criterion labeled by $v \in V$ can take values in a finite set $\mathcal{I}_v$. The resulting counts are gathered in a contingency table such that

$$\mathcal{I} = \prod_{v \in V} \mathcal{I}_v$$

is the set of cells $i = (v \in V)$. The vector of cell counts is denoted $n = (n(i), i \in \mathcal{I})$ with corresponding mean $m(i) = E(n) = (m(i), i \in \mathcal{I})$. For $D \subset V$,

$$\mathcal{I}_D = \prod_{v \in D} \mathcal{I}_v$$

is the set of cells $i_D = (i_v, v \in D)$ in the D-marginal table. The marginal counts are $n(i_D) = \sum_{j:j_D = i_D} n(j)$ with $m(i_D) = E\left(n(i_D)\right)$.

We assume that the components of n are independent and follow a Poisson distribution (i.e. Poisson sampling) and that the cell means are modeled according to a hierarchical model

$$\log(m) = X\theta$$

where X is an $|\mathcal{I}| \times p$ design matrix with rows $\{f_i, i \in \mathcal{I}\}$ and θ is a p-vector of log-linear parameters with $\theta \in R^p$. The results herein also apply under multinomial or product multinomial sampling. We assume that the first component of f_i is 1 for all $i \in \mathcal{I}$ and that "baseline constraints" are used making the $f_i's$ binary 0/1 vectors. For each cell, $i \in \mathcal{I}$, we have $\log m(i) = \langle f_i, \theta \rangle$.

The sufficient statistic $t = X^T n$ has the probability distribution in the natural exponential famil

$$f(t) = \exp\left(\langle \theta, t \rangle - \sum_{i \in \mathcal{I}} \exp\left(\langle f_i, \theta \rangle\right) \right) \nu\left(dt\right)$$

with respect to a discrete measure ν that has convex support

$$C_p = \left\{ \sum_{i \in \mathcal{I}} y(i) f_i, y(i) \geq 0, i \in \mathcal{I} \right\} = \text{cone}\left\{x_i, i \in \mathcal{I}\right\}$$

i.e. the convex cone generated by the rows of the design matrix X. The log-likelihood, as a function of m, is

$$l(m) = \sum_{i \in \mathcal{I}} n(i) \log m(i) - \sum_{i \in \mathcal{I}} m(i)$$

Let $\mathcal{M}$ be the column space of X. It is well known that the log-likelihood is strictly concave with a unique maximizer $\hat{m} = \text{argsup}_{\log m \in \mathcal{M}} l(m)$ that satisfies $X^T \hat{m} = t$.

Definition 2.1. If $\hat{m} > 0$, we say that $\hat{m}$ is the MLE of m while if $\hat{m}(i) = 0$ for some $i \in \mathcal{I}$ we call $\hat{m}$ the extended MLE.

The following important theorem from Haberman (1974) gives necessary and sufficient conditions for $\hat{m} > 0$, i.e. the existence of the MLE.

Theorem 2.2. *The MLE exists if and only if there exists a y such that $X^T y = 0$ and $y + n > 0$.*

Proof. Suppose that $\hat{m}$ exists. Then $X^T\hat{m} = t$ and $X^T(\hat{m} - n) = 0$. Letting $y = \hat{m} - n$ we have $X^Ty = 0$ and $y + n = \hat{m} > 0$. □

Conversely, suppose that there exists a y such that $X^Ty = 0$ and $y+n > 0$. Then $\sum_{i\in\mathcal{I}} y(i)\log m(i) = \sum_{i\in\mathcal{I}} y(i)\left\langle f_i, \theta\right\rangle = \left\langle \sum_{i\in\mathcal{I}} y(i)f_i, \theta\right\rangle = \left\langle X^Ty, \theta\right\rangle = 0$. We can then write the log-likelihood as

$$\begin{aligned} l(m) &= \sum_{i\in\mathcal{I}} n(i)\log m(i) - \sum_{i\in\mathcal{I}} m(i) \\ &= \sum_{i\in\mathcal{I}} (y(i) + n(i))\log m(i) - \sum_{i\in\mathcal{I}} m(i) \end{aligned}$$

Let $\mu = \log m$ and consider the real valued function $f\left(\mu(i)\right) = \left(y(i) + n(i)\right)\mu(i) - \exp\left(\mu(i)\right)$ for some $i \in \mathcal{I}$. Differentiating with respect to $\mu(i)$, we have $f'\left(\mu(i)\right) = \left(y(i) + n(i)\right) - \exp\left(\mu(i)\right)$ $f''(\mu(i)) = -\exp\left(\mu(i)\right) < 0$, and we see that f is strictly concave with a finite maximum $\mu(i) = \log\left(y(i) + n(i)\right)$ and l is bounded above. l is not bounded below, however, since $\lim_{\mu(i)\to\pm\infty} f\left(\mu(i)\right) = -\infty$.

Let A be the set $\left\{\mu \in \mathcal{M} : l\left(\mu\right) \geq c\right\}$ where $c \in R$. The number c can be chosen small enough such that the set A is non-empty. Then A is bounded and, since l is continuous function μ, it is closed. It follows that A is compact and l must have a finite maximum, $\hat{\mu}$ for some $\mu \in A$. We conclude that $\hat{m} > 0$.

Corollary 2.3. *If $n > 0$, the MLE exists.*

Proof. Take $y = 0$ in Theorem 2. □

3 Some basics of convex geometry

In this section we give some basic definitions that we need from convex geometry. Some supplementary references are Rockafellar (1970) and Ziegler (1995). In general, a polytope is a closed object with flat sides. The relative interior of a polytope is its interior with respect to the affine space of smallest dimension containing it. For a polytope that is full dimensional, the relative interior corresponds the the topological interior (int). The (convex) cone, generated by the points $a_1, a_2, ..., a_n$ is a polytope given by

$$\text{cone}\left\{a_1, a_2, ..., a_n\right\} = \left\{\sum_{i=1}^{n} a_i x_i : x_i \geq 0, i = 1, 2, ..., n\right\}$$

and its relative interior is

$$\left\{\sum_{i=1}^{n} \lambda_i a_i : \lambda_i > 0, i = 1, 2, ..., n\right\}$$

while the convex hull of the same points, $\text{conv}\left\{a_1, a_2, ..., a_n\right\}$, is also a polytope with the added restriction that $\sum_{i=1}^{n} \lambda_i = 1$. A convex polytope P can be represented as the convex hull of a finite number of points (the V-representation) or, equivalently, as the intersection of a finite number of half space (the H-representation). Cones and convex hulls are examples of convex polytopes. Henceforth, we assume that P is a convex polytope in R^d.

A face of P is a nonempty set of the form $F = P \cap \left\{x \in R^d : c^Tx = r\right\}$ where $c^Tx \leq r$ for all $x \in P$. The set $\left\{x \in R^d : c^Tx = r\right\}$ is called a supporting hyperplane to P. The faces of dimension 0 are called extreme points and, if P is a cone, the one dimensional faces of P are called the extreme rays of P. Moreover, when P is a cone all faces include the origin so that $r = 0$ and the origin is the only face of dimension 0. The dimension of a face F is the dimension of its affine hull

$$\text{aff}\left\{\sum_i \lambda_i x_i : x_i \in F, \sum_i \lambda_i = 1\right\}$$

which is the set of all affine combinations of the points in F. Finally, note that by taking $c = 0$, P itself is a face. We now have the following sequence of theorems. .

Theorem 3.1. *Any face of C^p of dimension at least one is the cone generated by the $f_i^\prime s$ that belong to that face.*

Proof. Suppose that t belongs to a face $F = C^p \cap \left\{x \in R^J : c^T x = 0\right\}$ of C_p of dimension at least one. Then F contains at least one point other than the origin. Let $\mathcal{I}_F = \{i \in \mathcal{I} : f_i \in F\}$. Every point in C^p can be expressed as a conical combination of the $f_i^\prime s$ and, hence, there exist non-negative real numbers $(\lambda_i, i \in \mathcal{I})$ such that $t = \sum_{i\in\mathcal{I}} \lambda_i f_i = \sum_{i\in\mathcal{I}_F} \lambda_i f_i + \sum_{i\in\mathcal{I}\backslash\mathcal{I}_F} \lambda_i f_i$. If $t = 0$, then since the first coordinate of f_i is 1, we must have $\lambda_i = 0$ for all $i \in \mathcal{I}$ and we can certainly write $t = \sum_{i\in\mathcal{I}_F} \lambda_i f_i$. If $t \neq 0$ then there must be an $i \in \mathcal{I}$ such that $\lambda_i > 0$. Suppose that $\lambda_i > 0$ for some $i \in \mathcal{I}\backslash\mathcal{I}_F$. Then

$$\begin{aligned} 0 &= c^T t = c^T \left(\sum_{i\in\mathcal{I}_F} \lambda_i f_i + \sum_{i\in\mathcal{I}\backslash\mathcal{I}_F} \lambda_i f_i\right) \\ &= \sum_{i\in\mathcal{I}_F} \lambda_i \left(c^T f_i\right) + \sum_{i\in\mathcal{I}\backslash\mathcal{I}_F} \lambda_i \left(c^T f_i\right) \\ &= \sum_{i\in\mathcal{I}\backslash\mathcal{I}_F} \lambda_i \left(c^T f_i\right) \\ &< 0 \end{aligned}$$

and we have a contradiction. Therefore, any $\lambda_i = 0$ for all $i \in \mathcal{I}\backslash\mathcal{I}_F$ and $t = \sum_{i\in\mathcal{I}_F} \lambda_i f_i$ which also implies that $\mathcal{I}_F \neq \emptyset$.

We have just shown that if $t \in F$ then t can be written as a conical combination of the $f_i^\prime s$ in $\mathcal{I}_F$. Let us show the converse. Indeed, for any set of non-negative real numbers $(\lambda_i, i \in \mathcal{I})$ with $\sum_{i\in\mathcal{I}_F} \lambda_i f_i \in F$ we have $c^T \left(\sum_{i\in\mathcal{I}_F} \lambda_i f_i\right) = \sum_{i\in\mathcal{I}_F} \lambda_i \left(c^T f_i\right) = 0$. □

A simple corollary of Theorem 2 is that there is always one and only one face of C^p that contains t in its relative interior, provided that $t \neq 0$. Let us not this formally.

Corollary 3.2. *If $t \in C^p$ and $t \neq 0$ then there is a unique face of C^p containing t in its relative interior.*

The next theorem pertains to determining the dimension of a face F, which is the dimension of aff (F). Henceforth, for a given face F of C^p we define$\mathcal{I}_F = \{i \in \mathcal{I} : f_i \in F\}$.

Theorem 3.3. *If F is a face of C_p of dimension at least one then* $\text{aff}(F) = \text{span}\{f_i, i \in \mathcal{I}_F\}$.

Proof. If $x \in \text{aff}\{F\}$ then there exist real numbers $\lambda_1, \lambda_2, ..., \lambda_k$ and points $x_1, x_2, ..., x_k \in F$ such that $\sum_{j=1}^k \lambda_j = 1$ and $x = \sum_{j=1}^k \lambda_j x_j$. Since $x_j \in F$ then $x_j = \sum_{i\in\mathcal{I}_F} \alpha_{ij} f_i$ for some non-negative real numbers $\alpha_{ij}, i \in \mathcal{I}_F$. Therefore, $x = \sum_{j=1}^k \lambda_j \left(\sum_{i\in\mathcal{I}_F} \alpha_{ij} f_i\right) \in \text{span}\{f_i, i \in \mathcal{I}_F\}$. Conversely, if

$x \in \operatorname{span}\{f_i, i \in \mathcal{I}_F\}$ then $x = \sum_{i \in \mathcal{I}_F} \lambda_i f_i$ for some real numbers $\lambda_i, i \in \mathcal{I}_F$. Since $0 \in F$ we can write

$$x = \left(1 - \sum_{i \in \mathcal{I}_F} \lambda_i\right) 0 + \sum_{i \in \mathcal{I}_F} \lambda_i f_i$$

which is an affine combination of points in F. Therefore, $x \in \operatorname{aff}(F)$.

Since there are p linearly independent $f_i's$, it follows that the dimension of C^p is p. □

Theorem 3.4. *If F is a face of C_p of dimension at least one then the extreme rays of F are the $f_i's$ that belong to that face.*

Proof. Suppose that $x \in F$ which implies that $x = \sum_{i \in \mathcal{I}_F} \lambda_i f_i$ for some non-negative real numbers $\lambda_i, i \in \mathcal{I}_F$ with at least one $\lambda_i > 0$. If x is an extreme ray then we must have exactly one $\lambda_i > 0$. For if not, then x would be a conical combination of two linearly independent vectors (none of the $f_i's$ are scalar multiples of one another). But then $x = \lambda_i f_i$.

For a given, $f_j \in F$ we need to show that f_j is an extreme ray. Suppose that this is not the case. Then f_j can be written as a conic combination of two vectors $x_1, x_2 \in C^p$ where $x_1 \neq kx_2$for some $k > 0$. That is, $f_j = \lambda_1 x_1 + \lambda_2 x_2$ for some $\lambda_1 \lambda_2 > 0$. But, by Theorem 2, $x_1 = \sum_{i \in \mathcal{I}_F} \alpha_{i1} f_i$ and $x_2 = \sum_{i \in \mathcal{I}_F} \alpha_{i2} f_i$ so that

$$f_j = \lambda_1 \sum_{i \in \mathcal{I}_F} \alpha_{i1} f_i + \lambda_2 \sum_{i \in \mathcal{I}_F} \alpha_{i2} f_i$$

Recalling that all the $f_i's$ are distinct binary 0/1 vectors we must have a contradiction since $\mathcal{I}_F \supseteq \{f_j\}$. □

The following corollary of Theorem 2 is due to Feinberg and Rinaldo (2012).

Corollary 3.5. *The MLE exists if and only if $t \in \operatorname{ri}(C^p)$.*

Proof. Suppose that the MLE exists. Then by Theorem 2, there exists a y such that $X^T y = 0$ and $y + n > 0$. But then $t = X^T(y + n)$ where $y + n > 0$ and hence $t \in \operatorname{ri}(C^p)$. Now suppose $t \in \operatorname{ri}(C^p)$. By definition, there exists an $y > 0$ such that $X^T y = t = X^T n$. But then $X^T(y - n) = 0$ and $n + (y - n) > 0$ and the MLE exists (by Theorem 2). □

4 An algorithm to determine $\mathcal{I}_F$

We have seen in Section 3, in particular, corollary (5), that there is a unique face F of C^p containing the sufficient statistic $t = X^T n$ in its relative interior. We turn now to finding F; for the MLE exists if and only if $F = C^p$. With $\mathcal{I}_F = \{i \in \mathcal{I} : f_i \in F\}$ we let X_F be an $\mathcal{I}_F \times J$ matrix with rows $f_i, i \in \mathcal{I}_F$. By Theorem 4, we know that $F = \operatorname{cone}\{f_i, i \in \mathcal{I}_F\}$ and by Theorem 6, the dimension of F is $p_F = \operatorname{rank}(X_F)$. Equipped with the following theorem, we can take an approach similar to Geyer (2009) and Feinberg and Rinaldo (2012), and finds $\mathcal{I}_F$ by solving a sequence of linear programs.

Theorem 4.1. *Any $a \geq 0$ in $R^{\mathcal{I}}$ such that $t = X^T a$ must have $a(i) = 0$ for any $i \in \mathcal{I} \backslash \mathcal{I}_F$.*

Proof. Since F is a face of C^p then it is of the form $F = C^p \cap \{x \in R^p : c^T x = 0\}$ where $c^T x < 0$ for $x \in C^p \backslash F$. Suppose that $t = X^T a = \sum_{i \in \mathcal{I}} a(i) f_i$ and $a(i) > 0$ for some $i \in \mathcal{I}_F$. Then

$$\begin{aligned} 0 &= \sum_{i \in \mathcal{I}_F} a(i) \left(c^T f_i\right) + \sum_{i \in \mathcal{I} \backslash \mathcal{I}_F} a(i) \left(c^T f_i\right) \\ &= \sum_{i \in \mathcal{I}_F} a(i) \left(c^T f_i\right) \\ &< 0 \end{aligned}$$

which is a contradiction. □

We now present an algorithm to find $\mathcal{I}_F$, which we call the facial set, and show that it works. The algorithm requires solving a sequence of linear programs that get progressively simpler until the problem is solved. Let $\mathcal{I}_0 = \{i \in \mathcal{I} : n(i) = 0\}$ and $\mathcal{I}_+ = \{i \in \mathcal{I} : n(i) > 0\}$. Before we begin, note that Theorem 9 applies to n and $\hat{m}$ since $X^T \hat{m} = X^T n = t$ so that $\mathcal{I}_+ \subseteq \mathcal{I}_F = \{i \in \mathcal{I} : \hat{m} > 0\}$ or, in other words, $n(i) > 0 \Rightarrow i \in \mathcal{I}_F$ and $i \in \mathcal{I}_F \iff \hat{m}(i) > 0$.

Algorithm 4.2 (A repeated linear programming algorithm to find $\mathcal{I}_F$)**.**

Input: The sufficient statistic t
Output: The facial set $\mathcal{I}_F$.

1. Set $A = \mathcal{I}_0$. If A is empty then set $\mathcal{I}_F = \mathcal{I} \backslash A = \mathcal{I}$. STOP

2. Solve the linear program (LP)

$$\begin{aligned} \max \quad & z = \sum_{i \in A} a(i) \\ s.t. \quad & X^T a = t \\ & a \geq 0 \end{aligned} \tag{4.1}$$

3. If the optimal objective value is $z = 0$ then set $\mathcal{I}_F = \mathcal{I} \backslash A$. STOP.

4. Let a be a feasible solution to the LP. For any $i \in A$ such that $a(i) > 0$ remove that index from A. Repeat for all feasbile solutions available (or even just the optimal solution).

5. If A is empty, then $\mathcal{I}_F = \mathcal{I} \backslash A = \mathcal{I}$. *STOP*. Otherwise return to STEP 2.

Keeping Theorem 9 in mind, let us now show that the algorithm works. If A is empty to being with then it is clear that $t = X^T n \in \text{ri}\left(C^p\right), F = C^p$, and $\mathcal{I}_F = \mathcal{I} \backslash A = \mathcal{I}$. The algorithm terminates at STEP 1. Suppose that A is not empty to start. Observe that, at any given iteration, the set $\mathcal{I} \backslash A$ is the set of $i \in \mathcal{I}$ such that we have found some $a \geq 0$ in $R^{\mathcal{I}}$ with $X^T a = t$ that has $a(i) > 0$. Next, observe that the algorithm terminates if A is empty or A is nonempty and the optimal objective value is $z = 0$. If A is empty then for all $i \in \mathcal{I}$ we have found some $a \geq 0$ in $R^{\mathcal{I}}$ with $x^T a = t$ that has $a(i) > 0$. It must be that $\mathcal{I}_F = \mathcal{I}$. If A is non-empty and the optimal objective value is $z = 0$ then every $a \geq 0$ such that $X^T a = t$ has $a(i) = 0$ for $i \in A$. The result is that $A = \mathcal{I} \backslash \mathcal{I}_F$ and $\mathcal{I}_F = \mathcal{I} \backslash A$. In all cases we have found $\mathcal{I}_F$.

Now, suppose that $y \in R^{\mathcal{I}}$ such that $y(i) > 0 \iff n(i) > 0$. Then $t' = X^T y = \sum_{i \in \mathcal{I}} y(i) f_i$ and $t = X^T n = \sum_{i \in \mathcal{I}} n(i) f_i$ belong to the same face F. Therefore, it is the location of the zero cells in n that determines F as opposed to the magnitude of the nonzero entries. For this reason, it is simplest to let s

$$y(i) = \begin{cases} 1 & n(i) > 0 \\ 0 & n(i) = 0 \end{cases}$$

and find $\mathcal{I}_F$ using t'.

For models where m is Markov with respect to a decomposable graph G the MLE (or extended MLE) a closed form expression for its computation exists. Theoretically, for such models, one need not resort to linear programming to find $\mathcal{I}_F$, since $\hat{m}$ can be computed exactly. Once this is done we know that $\mathcal{I}_F = \{i \in \mathcal{I} : \hat{m}(i) > 0\}$. Practically speaking though, it is simpler to use algorithm 10 for any model without first determining decomposability.

We now give an example applying Algorithm 10.

Example 4.3. Consider a $3 \times 3 \times 3$ table for variables a, b, c with counts:

0	1	0	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1
1	1	1	1	0	0	1	0	0

and the model $[ab][bc][ac]$. Suppose $\mathcal{I}_a = \mathcal{I}_b = \mathcal{I}_c = \{1, 2, 3\}$. The leftmost array corresponds to $c = 1$, the middle to $c = 2$, and the rightmost to $c = 3$. Let us apply our the algorithm to find F for this data set. We begin, at STEP 1, by setting

$$A = \mathcal{I}_0 = \{111, 131, 211, 322, 332, 323\}$$

where, by an abuse of notation, 131, refers to cell $i = (1, 3, 1)$ with $a = 1, b = 3, c = 1$. Since A is nonempty we proceed to STEP 2. The optimal solution to the (1) has $a(131) > 0$ and so we remove the cell 131 from A to get

$$A = \{111, 211, 322, 332, 323, 333\}$$

and return to STEP 2. Resolving (1) we find that, this time, the optimal objective value is 0. At this point we set $\mathcal{I}_F = \mathcal{I} \backslash A = \mathcal{I}_+ \cup \{131\}$ and the algorithm terminates. The dimension of F is $\operatorname{rank}(X_F)$, which in this case, is 18.

In the remainder of this section, we proceed to show that when $F \subset C$, maximum likelihood estimation can proceed almost as usual conditional on $t \in F$. Recall once more, the likelihood as a function of m.

$$\begin{aligned} L(m) &= \prod_{i \in \mathcal{I}} \exp\left(-m(i)\right) m(i)^{n(i)} \\ &= \prod_{i \in \mathcal{I}_F} \exp\left(-m(i)\right) m(i)^{n(i)} \prod_{i \in \mathcal{I} \backslash \mathcal{I}_F} \exp\left(-m(i)\right) m(i)^{n(i)} \end{aligned}$$

Since $n(i) = 0$ for $i \in \mathcal{I} \backslash \mathcal{I}_F$, then

$$\begin{aligned} L(m) &= \prod_{i \in \mathcal{I}_F} \exp\left(-m(i)\right) m(i)^{n(i)} \\ &= \exp\left(\sum_{i \in \mathcal{I}_F} n(i) \log m(i) - \sum_{i \in \mathcal{I}_F} m(i) \right) \\ &= \exp\left(\langle n_F, \log m_F \rangle - \sum_{i \in \mathcal{I}_F} m(i) \right) \end{aligned}$$

where $n_F = (n(i), i \in \mathcal{I}_F)$ and $m_F = (m(i), i \in \mathcal{I}_F)$. Let $\mathcal{M}_F$ be the linear span of the columns of X_F. Then $\hat{m}$ satisfies $\hat{m}(i) = 0$, $i \in \mathcal{I} \backslash \mathcal{I}_F$ and

$$\hat{m}_F = (\hat{m}(i), i \in \mathcal{I}_F) = argsup_{\log m_F \in \mathcal{M}_F} \exp\left(\langle n_F, \log m_F \rangle - \sum_{i \in \mathcal{I}_F} m(i) \right) \tag{4.2}$$

The conditional density of n given $t \in F$ is

$$\begin{aligned} P(n(i), i \in \mathcal{I} | t \in F) &= P(n(i), i \in \mathcal{I} | n(i) = 0, i \in \mathcal{I} \backslash \mathcal{I}_F) \\ &= P(n(i), i \in \mathcal{I}_F) \\ &= \prod_{i \in \mathcal{I}_F} \exp\left(-m(i)\right) m(i)^{n(i)} \\ &= \exp\left(\langle n_F, \log m_F \rangle - \sum_{i \in \mathcal{I}_F} m(i) \right) \end{aligned}$$

which is the same as (4.2). Therefore, when $F \subset C$, the MLE of m can be computed by conditional on $t \in F$. Practically speaking, this means that we can treat the zeros in the cells $i \in \mathcal{I} \backslash \mathcal{I}_F$ as structural zeros rather than sampling zeros. Now, if $t \in F$ and $d_F = \text{rank}(X_F) < d$ then X_F is not of full rank and the model $\log m_F = X_F \theta$ is over-parametrized; only d_F log-linear parameters will have finite estimates. We can partially fit the model by selecting d_F linearly independent columns of X_F, constructing a new design matrix X^*_F, and fitting the model $\log m_F = X^*_F \theta_F$. The new parameter vector θ_F will contain d_F components of θ which can be estimated. Estimates and standard errors of m_F and θ_F can then be obtained as usual. When the contingency table is not too sparse, and large sample χ^2 goodness of fit statistics are appropriate, the correct degrees of freedom is $|\mathcal{I}_F| - d_F$ (Feinberg and Rinaldo (2012)). It is an open research question whether the Bayesian Information Criterion Schwarz (1978) for comparing models should be corrected from $\hat{l} - \frac{d}{2} \log N$ to $\hat{l} - \frac{d_F}{2} \log N$ when $F \subset C$.

5 The eMLEloglin package

The main virtue of the eMLEloglin package is the ability to compute the facial set F for a given log-linear model and data set. It does this using algorithm 4.2 described above. The required linear programs are solved using the lpSolveAPI R package. If $F \subset C$, then a modified contingency table can be constructed, where cells in $\mathcal{I} \backslash \mathcal{I}_F$ are deleted, and passed to the GLM package to obtain maximum likelihood estimates. The GLM package will automatically identify a subset of the parameters that can be estimated and the correct model dimension.

The eMLEloglin package includes a sparse dataset from the household study at Rochdale referred to in Whittaker (1990). The Rochdale data set is a contingency table representing the cross classification of 665 individuals according to 8 binary variables. The study was conducted to elicit information about factors affecting the pattern of economic life in Rochdale, England. The variables are as follows: a. wife economically active (no, yes); b. age of wife >38 (no, yes); c. husband unemployed (no, yes); d. child≤ 4 (no, yes); e. wife's education, highschool+ (no, yes); f. husband's education, highschool+ (no, yes); g. Asian origin (no, yes); h. other household member working (no, yes). The table is sparse have 165 counts of zero, 217 counts with at most three observations, but also a few large counts with 30 or more observations.

Example 5.1. Consider the following $2 \times 2 \times 2$ contingency table for variables a, b, c with counts:

0	1	1	1
1	1	1	0

and the model $[ab][bc][ac]$. Suppose $\mathcal{I}_a = \mathcal{I}_b = \mathcal{I}_c = \{1, 2\}$. The leftmost array corresponds to $c = 1$, and the rightmost array to $c = 2$. This is one of the earliest known examples identified where the MLE does not exist Haberman (1974). We now show how to compute F using the eMLEloglin package. We first create a contingency table to hold the data:

```
> x <- matrix(nrow = 8, ncol = 4)
> x[,1] <- c(0,0,0,0,1,1,1,1)
> x[,2] <- c(0,0,1,1,0,0,1,1)
> x[,3] <- c(0,1,0,1,0,1,0,1)
> x[,4] <- c(0,1,2,1,4,1,3,0)
> colnames(x) = c("a", "b", "c", "freq")
> x <- as.data.frame(x, row.names = rep(,8))
>
> x a b c freq
  1 0 0 0 0
  2 0 0 1 1
  3 0 1 0 2
  4 0 1 1 1
  5 1 0 0 4
  6 1 0 1 1
  7 1 1 0 3
  8 1 1 1 0
```

We can then use the facial_set function:

```
> f <- facial_set (data = x, formula = freq ˜ a*b + a*c + b*c)
> f
$formula
freq ˜ a*b + a*c + b*c

$model.dimension
[1] 7

$status
```

```
[1] "Optimal objective value 0"

$iterations
[1] 1

$face
  a b c freq facial_set
1 0 0 0    0          0
2 0 0 1    1          1
3 0 1 0    2          1
4 0 1 1    1          1
5 1 0 0    4          1
6 1 0 1    1          1
7 1 1 0    3          1
8 1 1 1    0          0

$face.dimension
[1] 6

$maxloglik [1]
-1.772691
```

The output begins by giving the model formula and the original dimension. Under Poisson sampling the model of no three-way interaction has 7 free parameters. The line mentioning status is for debugging purposes to know how the algorithm terminated. For this example, Algorithm 4.2 terminated when an optimal objective value of $z = 0$ was found. The next line indicates that the algorithm required only one iteration to find F. The table in f2$face is probably the most important output. It indicates that

$$\begin{aligned} \mathcal{I}_F &= \{001, 010, 011, 100, 101, 110\} \\ \mathcal{I}\backslash\mathcal{I}_F &= \{000, 111\} \end{aligned}$$

The implication of the fact that, here, $\mathcal{I}_F \neq \mathcal{I}$ is that the dimension of F is 6 which we see under $face.dimension. Since $|\mathcal{I}_F| = d_F = 6$, the model is effectively saturated and the fitted values are the same as the observed values. We can see this by passing the data with the cells in $\mathcal{I}\backslash\mathcal{I}_F$ removed to the glm function.

```
> fit <- glm (formula = freq ~ a * b + a * c+ b * c,
              data = x[as.logical(f2$face$facial_set),]

> fit Call:  glm(formula = freq ~ a * b + a * c + b * c,
                 data = x[as.logical(f2$face$facial_set),      ])

Coefficients: (Intercept)            a            b            c
                2.000e+00    2.000e+00   -1.479e-15   -1.000e+00
        a:b          a:c        b:c
 -1.000e+00   -2.000e+00         NA
```

```
Degrees of Freedom: 5  Total (i.e. Null);  0 Residual
Null Deviance:   8  Residual Deviance: 6.015e-30
AIC: -383.4

> fit$fitted.values
2 3 4 5 6 7
1 2 1 4 1 3
```

As expected, one parameter is not able to be estimated; and R handles this automatically. Note that the residual degrees of freedom is correctly calculated to be $|\mathcal{I}_F| = d_F = 0$. Let us work through a larger example now with the Rochdale data.

Example 5.2. The Rochdale data comes preloaded with the package. Suppose we are interested in the model |ad|ae|be|ce|ed|acg|dg|fg|bdh| which is the model with the highest corrected BIC for this data set. We give a list of the top models by corrected and uncorrected BIC below for this data set. The required R code to find the facial set for this model is:

```
data(rochdale)
f <- facial_set  (data = rochdale,
                  formula = freq ~ a*d + a*e + b*e + c*e + e*f +
                                 a*c*g + d*g + f*g + b*d*h)
```

From the output we see that the model lies on a face of dimension 22. Since the original model dimension is 24, two parameters will not be estimable. Given the sparsity of the table, a goodness of fit test would not be appropriate. The fitted model can be obtained from the GLM function with the code:

```
fit <- glm (formula = freq ~ a*d + a*e + b*e + c*e + e*f +
                             a*c*g + d*g + f*g + b*d*h)
               data = rochdale[as.logical(f2$face$facial_set),])
```

The GLM function automatically determines that θ_{acg} and θ_{bdh} can not be estimated; which is because the acg and bdh margins are both zero. The residual degrees of freedom is correctly calculated at $|\mathcal{I}_F| - d_F = 196 - 22 = 174$.

Since the Rochdale dataset seems to be of some interest recently we give the top five models in terms of corrected and the usual BIC (abbreviated cBIC and BIC in the tables, respectively).

	cBIC	Model Dim.	Face Dim.
|ad|ae|be|cd|ef|acg|dg|fg|bdh|	985.3	24	22
|ad|ae|be|ce|cf|ef|acg|dg|fg|bdh|	985.2	25	23
|ad|ae|be|ce|cf|df|ef|acg|dg|fg|bdh	984.4	26	24
|ad|ae|be|ce|df|ef|acg|dg|fg|bdh|	984.3	25	23
|ac|ad|ae|be|ce|ef|ag|cg|dg|fg|bdh	984.0	23	22

Model	BIC	Model Dim.	Face Dim.
|ac|ad|bd|ae|be|ce|ef|ag|cg|dg|fg|bh|dh|	981.3	22	22
|ac|ad|bd|ae|be|ce|cf|ef|ag|cg|dg|fg|bh|dh|	981.1*	23	23
|ac|ad|ae|be|ce|ef|ag|cg|dg|fg|bdh|	980.7	23	22
|ac|ad|ae|be|ce|cf|ef|ag|cg|dg|fg|bdh|	980.5**	24	23
|ac|ad|bd|ae|be|ce|ef|ag|cg|dg|fg|bh|	980.4	21	21

With the exception of cf and the three factor interactions, the model with the highest uncorrected BIC is the model identified by Whitakker who, in any case, limited himself to considering at most two-factor interactions because of the sparsity of the table. Whittaker fit the all two factor interaction model, and then deleted the terms that we non-significant and arrived at the model |ac|ad|bd|ae|be|ce|cf|ef|ag|cg|dg|fg|bh|dh marked by an asterisk (*) above.

We note that bdh interaction has also been identified by Dobra and Massam (2010). The model they selected is marked with (**) above. Using Mosaic plots, Hofmann (2003) also observed that there is a strong hint of bdh interaction. The dataset was also analyzed in Dobra and Lenkoski (2011).

References

A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, Hoboken, NJ, 2 edition, 1990.

Y. M. M. Bishop, S. E. Feinberg, and P. W. Holland. *Discrete Multivariate Analysis.* MIT Press, Cambridge, MA, 1975.

R. Christensen. *Log-linear Models and Logistic Regression.* Springer Verlag, 1997.

A. Dobra and A. Lenkoski. Copula gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5:969–993, 2011.

A. Dobra and H. Massam. The mode oriented stochastic search (moss) algorithm for log-linear models with conjugate priors. *Statistical Methodology*, 7:204–253, 2010.

N. Eriksson, S. E. Feinberg, A. Rinaldo, and S. Sullivant. Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computations*, 41:222–233, 2006.

S. E. Feinberg and A. Rinaldo. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137:3430–3445, 2007.

S. E. Feinberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40:996–1023, 2012.

C. J. Geyer. Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259–289, 2009.

S. J. Haberman. *The Anlaysis of Frequency Data.* University of Chicago Press, 1974.

H. Hofmann. Constructing and reading mosaicplots. *Computational Statistics and Data Analysis*, 43:565–580, 2003.

S. F. Lauritzen. *Graphical Models.* Oxford University Press, NY, 1996.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `http://www.R-project.org/`.

R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* John Wiley & Sons, Chichester, 1990.

M. G. Ziegler. *Lectures on Polytopes.* Springer-Verlag, NY, 1995.