Maximum Likelihood Estimation in Log-Linear Models Supplementary Material: Algorithms

Stephen E. Fienberg*
Department of Statistics
Heinz College
Machine Learning Department
Cylab
Carnegie Mellon University
Pittsburgh, PA 15213 USA

Alessandro Rinaldo†
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213 USA

1 Introduction

This document contains the supplementary material to the article "Maximum Likelihood Estimation in Log-Linear Models" by S.E. Fienberg and A. Rinaldo, which henceforth we will refer to as FR.

We use the theory developed in FR to derive efficient algorithms for extended maximum likelihood estimation in log-linear models under Poisson and product multinomial schemes. The restriction to these sampling schemes is motivated by a variety of reasons. First, these schemes encode sampling constraints that arise most frequently in practice. In particular, these are the sampling schemes practitioners use in fitting hierarchical log-linear models, and especially the class of graphical models. Second, for these particular sampling schemes the log-partition function has a closed form expression and we can easily optimize the associated log-likelihood. Finally, as shown in theorem 9 of FR, the extended MLE of the cell mean value is identical in the two sampling schemes and, for the product multinomial scheme, the estimator is in fact the conditional MLE of the cell means given the sample constraints. Thus these estimates are highly interpretable. Some of the algorithms described in this document are implemented in a `MATLAB` toolbox available at http://www.stat.cmu.edu/~arinaldo/ExtMLE/.

We begin with a high-level overview of extended maximum likelihood estimation, summarizing the theoretical contributions from the previous section and laying down the rationale for the algorithm we propose. To simplify the exposition, we initially develop our result for the simpler case of a Poisson sampling scheme, and later treat the more complex case of product multinomial schemes.

Consider a log-linear model with associated d-dimensional log-linear subspace $\mathcal{M}$ and design matrix A, which for simplicity we assume to be of full-rank d. (When A is not of full rank, we need only minor changes to the arguments.) We focus on the problem of estimating the cell mean values of the corresponding extended exponential family based on the observed table $\mathbf{n}$. From the results described in section 3.1 of FR, we know that the MLE of $\boldsymbol{\mu}$ and, therefore, of $\mathbf{m}$, exists if and only if the observed sufficient statistics $\mathbf{t} = \mathrm{A}^\top \mathbf{n}$ lie in the interior of the d-dimensional marginal cone C_{A}. In this case, the log-likelihood, parametrized either using the log-linear parameters $\boldsymbol{\mu} \in \mathcal{M}$ or the natural parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ is a concave function admitting a unique optimizer with finite norm, the maximum likelihood estimate. The MLE does not existent if and only if $\mathbf{t} \in \mathrm{ri}(F)$, for some face F of C_{A} of dimension $d_F < d$ with associated facial set $\mathcal{F}$. Notice that F, d_F and $\mathcal{F}$ are random, since they depend on $\mathbf{t}$. Nonexistence of the MLE implies non-estimability of

[*] Email: `fienberg@stat.cmu.edu`

[†] Email: `arinaldo@cmu.edu`

both the log-linear and natural parameters, as formalized in theorem 7 of FR. The log-likelihood function is still concave, though not strongly so (see remarks in section 3.1), and, under the natural parametrization, it contains directions of recession, given by the normal cone to the face of the marginal cone containing in its relative interior the observed sufficient statistics (see Rinaldo et al., 2009, corollary 2.8). While this is an issue that cannot be resolved unless more data become available or we consider a different model, of dimension no larger than d_F, the theory of extended exponential family provides the theoretical justification for identifying a subset of size d_F of the original parameters that are in fact estimable. For the log-linear parameters, we can construct this subset as follows. Let $\mathrm{A}_{\mathcal{F}}$ denote the matrix obtained by considering only the rows of A with coordinates in $\mathcal{F}$, so that $\mathrm{rank}(\mathrm{A}_{\mathcal{F}}) = d_F$. Then, the columns of $\mathrm{A}_{\mathcal{F}}$ span $\mathcal{M}_{\mathcal{F}}$, the d_F-dimensional linear subspace of $\mathbb{R}^{\mathcal{F}}$ obtained as the coordinate projection of $\mathcal{M}$. The set $\mathcal{M}_{\mathcal{F}}$ is the set of log-linear parameters for the extended exponential family describing the restriction of the log-linear model $\mathcal{M}$ to the cells in $\mathcal{F}$. Within this restricted family, the MLE of the log-linear parameter exists and is given by a unique point $\widehat{\boldsymbol{\mu}}_{\mathcal{F}} \in \mathcal{M}_{\mathcal{F}}$, with the corresponding MLE for the cell mean value given by $\widehat{\mathbf{m}}_{\mathcal{F}} = \exp(\widehat{\boldsymbol{\mu}}_{\mathcal{F}})$, so that $\tau_{\mathcal{F}}(\widehat{\mathbf{m}}_{\mathcal{F}}) \in \overline{M}$. For the natural parametrization of the restricted family, it is sufficient to replace the $|\mathcal{F}| \times d$ design matrix $\mathrm{A}_{\mathcal{F}}$, which is not of full-rank, with any another matrix $\mathrm{A}^*_{\mathcal{F}}$ of full rank d_F and identical column range, i.e., to use a minimal representation. Then the natural parameter space for the restricted model becomes $\mathbb{R}^{d_F}$.

Once we identify the random facial set $\mathcal{F}$ corresponding to the observed sufficient statistic $\mathbf{t}$, extended maximum likelihood estimation is a relatively straightforward problem from the computational standpoint. Indeed, under natural parametrization and using a full-rank design matrix, the extended log-likelihood is a strictly concave function on $\mathbb{R}^{d_F}$ with no direction of recessions, thus admitting a unique minimizer, which we can compute efficiently using Newton-Raphson procedure (see section 3 below). The computational difficulties in extended maximum likelihood estimation rest mainly in isolating the coordinates comprising $\mathcal{F}$. Due to the combinatorial complexity of the face lattice of C_{A}, facet enumeration is computationally infeasible, even for small models, such as those in the examples in section 4 of FR. Thus, we need algorithms for isolating $\mathcal{F}$ that are applicable to large tables and complex models.

The computational procedure we propose for maximum likelihood estimation proceeds in two fundamental steps, described in detail below and summarized in Table 1. The input to the procedure is the design matrix A and the observed table $\mathbf{n}$.

1. **Identification of the facial set (section 2).** Computing the facial set is a task that corresponds to:

   Given a conic integer combination $\mathbf{t} = \mathrm{A}^\top \mathbf{n}$ *of the columns of* A*, determine the set* $\mathcal{F}$ *of those columns which span the face of* C_{A} *containing* $\mathbf{t}$ *in its relative interior.*

   For this task, the design matrix does not have to be of full rank, even though this is preferable.

2. **Log-likelihood optimization (section 3).**
   After we obtain the appropriate facial set $\mathcal{F}$, if $\mathcal{F} = \mathcal{I}$, then the MLE exist and can be obtained as

   $$\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^d} \ell^P(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^d}\ \mathbf{t}^\top \boldsymbol{\theta} - \mathbf{1}^\top \exp(\mathrm{A}\boldsymbol{\theta}). \tag{1}$$

   We have slightly abused our notation by writing $\ell^P(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$ in lieu of $\ell^P(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \mathcal{M}$, as originally defined in equation (2) of RF. Since $\mathbb{R}^d$ and $\mathcal{M}$ are isomorphic, this is inconsequential. We can carry out the optimization of ℓ^P using the Newton-Raphson method, but A must be of full rank in order for ℓ^P to have a unique optimizer. The MLE of the cell mean vector is $\widehat{\mathbf{m}} = \exp(\mathrm{A}\widehat{\boldsymbol{\theta}})$.

   If $\mathcal{F} \subsetneq \mathcal{I}$, and thus t the MLE does not exist, a new, we can compute a reduced design matrix $\mathrm{A}^*_{\mathcal{F}}$ of rank d_F by selecting any subset of linearly independent rows from $\mathrm{A}_{\mathcal{F}}$, e.g., using in proposition 5.1 from section 5. The extended likelihood is strictly concave and admits a unique optimizer, the extended MLE:

   $$\widehat{\boldsymbol{\theta}}^{\mathrm{e}} = \mathrm{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{d_F}} \ell^P_{\mathcal{F}}(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{d_F}}\ \mathbf{t}_F^\top \boldsymbol{\theta} - \mathbf{1}^\top \exp(\mathrm{A}^*_{\mathcal{F}}\boldsymbol{\theta}), \tag{2}$$

   where $\mathbf{t}_F = (\mathrm{A}^*_{\mathcal{F}})^\top \mathbf{n}_{\mathcal{F}}$. Our primary approach uses the Newton-Raphson procedure, but we could substitute alternatives for specific purposes, as we note below. Note that we use only the observed

Table 1: Pseudo-code for extended maximum likelihood estimation under Poisson sampling.

```
Input:  $\mathrm{A}$ and $\mathbf{t}$

Identification of the facial set
  Compute $\mathcal{F}$
Log-likelihood optimization
  if $\mathcal{F} = \mathcal{I}$
      compute $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ as in (1)
      return $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{m}} = \exp(\mathrm{A}\widehat{\boldsymbol{\theta}})$
  else
      find $\mathrm{A}^*_{\mathcal{F}}$ such that $\mathcal{R}(\mathrm{A}^*_{\mathcal{F}}) = \mathcal{M}_{\mathcal{F}}$ and $\mathrm{rank}(\mathrm{A}^*_{\mathcal{F}}) = \dim(\mathcal{M}_{\mathcal{F}}) = d_F$
      compute $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{d_F}$ as in (2)
      return $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{m}} = \tau_{\mathcal{F}}\left(\exp(\mathrm{A}^*_{\mathcal{F}}\widehat{\boldsymbol{\theta}})\right)$
  end
```

counts corresponding to cells in $\mathcal{F}$ in the optimization of the extended likelihood. In fact, the mean values for the cells in $\mathcal{F}^c$, the *likelihood zeros*, are not estimable and, therefore, we set them to zero. The extended MLE of the cell mean vector is

$$\widehat{\mathbf{m}}^{\mathrm{e}} = \tau_{\mathcal{F}}\left(\exp(\mathrm{A}^*_{\mathcal{F}}\widehat{\boldsymbol{\theta}}^{\mathrm{e}})\right) \in \overline{M}.$$

As a by-product of this procedure, we obtain a basis for the subspace $\mathcal{M}_{\mathcal{F}}$, whose dimension is also the dimension of the boundary log-linear model, or the order of the reduced exponential family corresponding to $\mathcal{F}$.

We close this introductory section with a remark about detecting existence of the MLE. If we only need to decide whether the MLE exists or not, then it is sufficient to set up the following linear program:

$$\begin{array}{rl} & \max s \\ \text{s.t.} & \mathrm{A}\mathbf{x} = \mathbf{t} \\ & \mathbf{x}_i - s \geqslant 0, \quad \forall i \\ & s \geqslant 0. \end{array}$$

The MLE does not exists if and only if the optimum s^* is zero, because that implies that there does not exist any strictly positive vector $\mathbf{x}$ with $\mathbf{t} = \mathrm{A}\mathbf{x}$, which is equivalent to $\mathbf{t}$ lying on the boundary of C_{A}. Clearly, this procedure cannot be used to detect parameter estimability or evaluate the effective dimension of the model under a nonexistent MLE.

Maximum Likelihood Estimation and Cuts

Maximum likelihood estimation under general conditional Poisson sampling scheme is typically computationally intractable. Indeed, unless the log-partition function ψ has a known closed form, its evaluation requires summing over all the points in $S(\mathrm{V})$, a task that becomes computationally too expensive to carry out even in models of small dimension.

In the special case where $\mathrm{V}^\top \mathbf{n}$ is a cut (see Bardorff-Nielsen, 1978) for the exponential family arising from the unrestricted Poisson scheme, there exists a strategy for maximum likelihood estimation that is computationally feasible. In fact, suppose that A has the form given in equation (8) of RF, and partition the vector of natural parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ and of sufficient statistics $\mathbf{t} = \mathrm{A}^\top \mathbf{n} = (\mathbf{t}^{(1)}, \mathbf{t}^{(2)})$ accordingly, where $\boldsymbol{\theta}^{(1)}, \mathbf{t}^{(1)} \in \mathbb{R}^{d-m}$ and $\boldsymbol{\theta}^{(2)}, \mathbf{t}^{(2)} \in \mathbb{R}^{m}$. It follows from lemma 1 of RF and the subsequent remarks that $\mathbf{t}^{(1)}$ is the vector of sufficient statistics and that only $\boldsymbol{\theta}^{(1)}$ is estimable. In practice, one could always maximize the Poisson likelihood, using $(\mathbf{t}^{(1)}, \mathbf{1})$ as the sufficient statistics, where $\mathbf{1} \in \mathbb{R}^m$. Effectively, this is equivalent to disregarding the fact that some of the sufficient statistics, namely the entries of $\mathbf{t}^{(2)}$, are fixed by design, treating them as random instead, and optimizing the Poisson likelihood function with respect to $\boldsymbol{\theta}$, which is computationally tractable. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\boldsymbol{\theta}}^{(2)})$ denote the optimum value, assumed finite and partitioned in the fashion described above. Furthermore, let $\hat{\boldsymbol{\theta}}^{(1)}$ the actual MLE of $\boldsymbol{\theta}^{(1)}$, once again assumed it exists. Then, the arguments described on page 128 of Bardorff-Nielsen (1978) yield that, if $\mathbf{t}^{(2)}$ is a cut, $\hat{\boldsymbol{\theta}}^{(1)} = \tilde{\boldsymbol{\theta}}^{(1)}$. These results generalize to the extended maximum likelihood estimation. In the interest of space, we omit the details, but, as a concrete example, we point out that, when V is the sampling matrix encoding the product multinomial scheme constraints, it is easy to see that $\mathrm{V}^\top \mathbf{n}$ is a cut. Thus we can view theorem 7 of RF as a special case of this more general phenomenon.

2 Determination of the Facial Sets

We derive two methods for determining facial sets, one based on linear programming and the other on the maximization of a non-linear function via Newton-Raphson procedure. We describe alternative methodologies in section 6. Throughout this section, we denote with A_+ and A_0 the sub-matrices obtained from A by considering the rows indexed by $\mathcal{I}_+ := \mathrm{supp}(\mathbf{n})$ and $\mathcal{I}_0 := \mathrm{supp}(\mathbf{n})^c$, respectively.

Recall that each face F of the marginal cone C_{A} is uniquely identified by the associated facial set $\mathcal{F} \subset \mathcal{I}$, which is determined by the conditions that

$$\begin{cases} \mathbf{a}_i^\top \mathbf{c} = 0 & \text{if} \quad i \in \mathcal{F} \\ \mathbf{a}_i^\top \mathbf{c} > 0 & \text{if} \quad i \in \mathcal{F}^c, \end{cases} \tag{3}$$

where $\mathbf{a}_i$ denotes the i-th row of U and $-\mathbf{c}$ is any point in the interior of the normal cone to the face F corresponding to $\mathcal{F}$. For simplicity we call the set $\mathcal{F}^c = \mathcal{I} \backslash \mathcal{F}$ the *co-facial set* of F. Without loss of generality, we have switched the sign of the inequalities from the original definition given in equation (10) of RF.

Equation (3) implies that the observed sufficient statistics $\mathbf{t} = \mathrm{A}^\top \mathbf{n}$ belong to the relative interior of some proper face F of the marginal cone if and only if the associated co-facial set $\mathcal{F}^c$ satisfies the inclusion $\mathcal{F}^c \subseteq \mathcal{I}_0$. This, in turn, is equivalent to the existence of a vector $\mathbf{c}$ satisfying:

1. $\mathrm{A}_+ \mathbf{c} = \mathbf{0}$;
2. $\mathrm{A}_0 \mathbf{c} \gneqq \mathbf{0}$;
3. the set $\mathrm{supp}(\mathrm{A}\mathbf{c})$ has maximal cardinality among all sets of the form $\mathrm{supp}(\mathrm{A}\mathbf{x})$ with $\mathrm{A}\mathbf{x} \gneqq \mathbf{0}$.

Therefore, any solution $\mathbf{x}^*$ of the non-linear optimization problem

$$\begin{array}{ll} \max & |\mathrm{supp}(\mathrm{A}\mathbf{x})| \\ \text{s.t.} & \mathrm{A}_+ \mathbf{x} = 0 \\ & \mathrm{A}_0 \mathbf{x} \geqslant \mathbf{0} \end{array} \tag{4}$$

will identify the required co-facial set $\mathcal{F}^c = \mathrm{supp}(\mathrm{A}\mathbf{x}^*)$. In particular, the MLE exists if and only if $\mathcal{F}^c = \varnothing$. It is worth pointing out that the non-existence of the MLE only depends on the *location* of the sampling zeros, not on the magnitude of the non-zero cells. Thus we can simplify all calculations using a table of 0s and 1s.

The problem (4) can be simplified making use of the following, simple fact.

Lemma 2.1. *The MLE exists if* $\mathrm{rank}(\mathrm{A}_+) = \mathrm{rank}(\mathrm{A})$.

Proof. If $\mathrm{rank}(\mathrm{A}_+) = \mathrm{rank}(\mathrm{A})$, every row of A_0 is a linear combination of the rows of A_+. Thus, for any vector $\mathbf{c}$ with $\mathrm{A}_+\mathbf{c} = \mathbf{0}$, it must be that $\mathrm{A}_0\mathbf{c} = \mathbf{0}$ as well. This implies that the feasible set for the problem (4) is kernel(A), hence $\mathcal{F}^c = \varnothing$, so the MLE exits. ■

The condition of lemma 2.1 is only sufficient, as in the following example demonstrates.

Example 2.2. If $\mathrm{rank}(\mathrm{A}_+) < \mathrm{rank}(\mathrm{A})$, the MLE may still exist. Indeed, consider the 3-way table

0	0	
0		0

0	0	
0	0	

		0

where the empty cells correspond to positive counts. For the hierarchical model [12][13][23], the MLE is well-defined but $\mathrm{rank}(\mathrm{A}_+) = 18$ and $\mathrm{rank}(\mathrm{A}) = 19$. ■

In light of lemma 2.1, it is necessary to look for a facial set only when $\mathrm{rank}(\mathrm{A}) > \mathrm{rank}(\mathrm{A}_+)$. If this is in fact the case, define the matrix $\mathrm{B} = \mathrm{A}_0\mathrm{Z}$, where the columns of Z form a basis for $\mathrm{kernel}(\mathrm{A}_+)$. Note that $\mathrm{rank}(\mathrm{B}) = q$, with $q = \mathrm{codim}(\mathcal{R}(\mathrm{A}_+)) = d - \mathrm{rank}(\mathrm{A}_+)$ and that B is of full rank. Next, observe that (the permutation of the elements of) any vector $\mathbf{y} \in \mathcal{R}(\mathrm{A})$ with $\mathbf{y}_{\mathcal{I}_0} = \mathbf{0}$ can be written as

$$\mathbf{y} = \mathrm{A}\mathrm{Z}\mathbf{x} = \left(\begin{array}{c} \mathrm{A}_+\mathrm{Z}\mathbf{x} \\ \mathrm{A}_0\mathrm{Z}\mathbf{x} \end{array} \right) = \left(\begin{array}{c} \mathbf{0} \\ \mathrm{B}\mathbf{x} \end{array} \right),$$

for some $\mathbf{x} \in \mathbb{R}^q$. The nonzero rows of B are indexed in a natural way by the corresponding subset of $\mathcal{I}_0$, denoted with $\mathcal{I}_\mathrm{B}$. In the remainder of the section it is assumed, without loss of generality, that B does not have any zero rows, namely $\mathrm{B} = \mathrm{B}_{\mathcal{I}_\mathrm{B}}$. Then, another condition for existence of the MLE follows readily.

Corollary 2.3. *Consider the non-linear optimization problem*

$$\begin{array}{ll} \max & |\mathrm{supp}(\mathrm{B}\mathbf{x})| \\ \text{s.t.} & \mathrm{B}\mathbf{x} \geqslant 0. \end{array} \tag{5}$$

The MLE exists if and only if the system $\mathrm{B}\mathbf{x} \gneqq \mathbf{0}$ *is infeasible. Any optimal solution* $\mathbf{x}^*$ *of (5) will identify the co-facial set* $\mathcal{F}^c = \mathrm{supp}(\mathrm{B}\mathbf{x}^*)$

In order to compute the matrix B, we need to determine a basis for $\mathrm{kernel}(\mathrm{U}_+)$, if it is different than the trivial subspace $\{\mathbf{0}\}$, e.g., using the results discussed in section 5, and, in particular, equation (22).

We consider two methods for finding a solution to problem (5), one based on linear programming, the second one on non-linear optimization. See section 6 for alternative procedures.

2.1 Linear Programming

Although the optimization problem (5) is highly non-linear, we can still use linear programming (LP) methods to compute its solution. The non-linearity is in fact problematic to the extent that it typically requires repeated implementations of LP algorithms, whose complexity, however, decrease at each iteration.

A linear relaxation of problem (4) leads to the linear program

$$\begin{array}{ll} \max & \left(\mathbf{1}_0^\top \mathrm{A}_0\right)\mathbf{x} \\ \text{s.t.} & \mathrm{A}_+\mathbf{x} = 0 \\ & \mathrm{A}_0\mathbf{x} \geqslant \mathbf{0} \\ & \mathrm{A}_0\mathbf{x} \leqslant \mathbf{1}, \end{array} \tag{6}$$

where the third constraint is required to bound the value of the objective function. The feasible set contains kernel(A) and is contained in the dual cone of C_A. If $\mathbf{x} \in \mathrm{kernel}(\mathrm{A})$, the objective function takes on its maximum value 0. In fact, the MLE exists if and only if the feasible set reduces to kernel(A).

It is convenient to take advantage of the simplified problem (5) to re-formulate (6) more compactly as

$$\begin{array}{rccc} & \max \mathbf{1}^{\top}\mathbf{y} & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{B}\mathbf{x} \\ & \mathbf{y} & \geqslant & \mathbf{0} \\ & \mathbf{y} & \leqslant & \mathbf{1}. \end{array} \tag{7}$$

If $(\mathbf{x}^*, \mathbf{y}^*)$ is a pair of optimal solutions, $\mathbf{1}^{\top}\mathbf{y}^* = 0$ if and only if the MLE exist, which happens if and only if $\mathbf{0}$ is the only point in the feasible set. When the MLE does not exist, the optimal solution $\mathbf{y}^*$ is not necessarily the one with maximal support and, consequently, it would not identify the correct facial set, but instead a larger facial set corresponding to a face of C_{A} which contains $\mathbf{t}$ on its relative boundary. This is illustrated in the next example.

Example 2.4. For the case of 4^3 tables and the hierarchical log-linear model of no-3-factor effect [12][23][13], consider the following table

0	0	0	0
2	0	0	2
5	1	0	0
0	1	0	0

0	4	1	0
1	0	1	0
0	0	0	0
0	2	4	0

0	0	0	3
0	0	0	0
3	5	3	5
0	0	0	0

0	2	0	5
3	0	0	1
0	0	0	0
0	4	0	0

,

in which the zeros are likelihood zeros obtained by taking the union of two among the 113,740 possible patterns of likelihood zeros characterizing the facets of the corresponding marginal cone. Using the MATLAB routine linprog[1], one application of the LP procedure identifies only a subset of likelihood zeros, namely

	0		0
		0	0

			0
			0
		0	

	0		
0			0
	0		

,

and to correctly determine the complete pattern, we need to use a second iteration, after removing the likelihood zeros found in the first one. ■

The above example suggests that repeated applications of (7) will eventually produce the required co-facial set: replace B with $\mathrm{B}_{\mathrm{supp}(\mathrm{B}\mathbf{x}^*)^c}$, where $\mathrm{supp}(\mathrm{B}\mathbf{x}^*)^c = \mathcal{I}_{\mathrm{B}} \backslash \mathrm{supp}(\mathrm{B}\mathbf{x}^*)$ (so that $\mathcal{I}_{\mathrm{B}}$ becomes smaller) and iterate, until either the objective function is 0 or $\mathrm{supp}(\mathrm{B}\mathbf{x}^*)^c = \varnothing$. Table 2 provides the details of the algorithm, which consists of a sequence of linear programs of decreasing complexity, while the next result shows its correctness.

Lemma 2.5. *Let* $(\mathbf{x}^1, \mathbf{y}^1)$ *an the optimal solution for the first iteration of the algorithm described in Table* (2)*. Then* $\mathbf{1}^{\top}\mathbf{y}^1 = 0$ *if and only if the MLE exist. If the MLE does not exist, the algorithm will return* $\mathcal{F}$ *in a finite number of iterations.*

Proof. In the first round of the procedure, $\mathbf{1}^{\top}\mathbf{y}^1 > 0$ if and only if the system $\mathrm{B}\mathbf{x} \gneqq \mathbf{0}$ is feasible, which is equivalent to the existence of the MLE by corollary (2.3).

As for the second claim, suppose that the MLE does not exist, that is $\mathbf{t} \in \mathrm{ri}(F)$ for some face F of C_{A}. Then, let $(\mathbf{x}^k, \mathbf{y}^k)$ be the optimal solutions returned based on the coefficient matrix B_k at the k-th round of the algorithm, $k \geqslant 2$, so that $\mathrm{B} = \mathrm{B}_1$. Explicitly, B_k is the submatrix of B_{k-1} obtained by considering only the rows indexed by the coordinates $\{i \colon \mathbf{y}_i^{k-1} = 0\}$.

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

Table 2: Pseudo-code for the LP procedure to compute the facial set $\mathcal{F}$. Recall that the rows of B are indexed by the set $\mathcal{I}_{\mathrm{B}}$. At each round of the algorithm, only the rows of B with indexes in $\mathrm{supp}(\mathbf{y}^*) \subseteq \mathcal{I}_{\mathrm{B}}$ are retained.

```
$\mathcal{F}$ = $\mathcal{I}$
do repeat
  compute a solution $(\mathbf{y}^*, \mathbf{x}^*)$ of (7)
  if $\mathbf{1}^\top \mathbf{y}^*$ = $0$
      return $\mathcal{F}$
  else
      $\mathcal{F}$ = $\mathcal{F} \backslash \mathrm{supp}(\mathbf{y}^*)$
      if $\mathrm{supp}(\mathbf{y}^*)^c$ = $\varnothing$
          return $\mathcal{F}$
      else
          $\mathrm{B}$ = $\mathrm{B}_{\mathrm{supp}(\mathbf{y}^*)^c}$
      end
  end
end
```

Suppose that $\mathbf{1}^\top \mathbf{y}^k > 0$. Notice that, while, necessarily, $\mathbf{y}^k = \mathrm{B}_k \mathbf{x}_k \geqslant \mathbf{0}$, the coordinates of $\mathrm{B}\mathbf{x}^k$ are not guaranteed to be non-negative. Below, we will rescale $\mathbf{x}_k$ appropriately so that the resulting vector $\tilde{\mathbf{x}}_k$ is such that $\mathrm{supp}(\mathrm{B}_k \tilde{\mathbf{x}}_k) = \mathrm{supp}(\mathrm{B}_k \mathbf{x}_k)$ and, at the same time,

$$\mathrm{B}\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_j\right) \gneqq \mathbf{0}.$$

To this end, for a generic vector $\mathbf{w}$, let

$$\mathrm{smax}(\mathbf{w}) = \max\left\{|\mathbf{w}_i| \colon i \in \mathrm{supp}(\mathbf{w})\right\} \quad \text{and} \quad \mathrm{smin}(\mathbf{w}) = \min\left\{|\mathbf{w}_i| \colon i \in \mathrm{supp}(\mathbf{w})\right\}.$$

Next, set $\tilde{\mathbf{x}}^1 = \mathbf{x}^1$ and

$$\tilde{\mathbf{x}}_k = \frac{1}{\mathrm{smax}(\mathrm{B}\mathbf{x}_k)} \frac{\mathrm{smin}(\mathrm{B}\tilde{\mathbf{x}}_{k-1})}{2} \mathbf{x}^k, \quad k > 1.$$

Because the entries of $\tilde{\mathbf{x}}_k$ are proportional to the entries to $\mathbf{x}_k$, is clear that $\mathrm{supp}(\mathrm{B}_k \tilde{\mathbf{x}}_k) = \mathrm{supp}(\mathrm{B}_k \mathbf{x}_k)$. Furthermore, due to the recursive nature of the normalizations described above, we also obtain

$$\mathrm{B}\mathbf{z}_k \gneqq \mathbf{0}.$$

where

$$\mathbf{z}_k = \sum_{j=‘}^{k} \tilde{\mathbf{x}}_j,$$

as claimed. Thus, $-\mathbf{z}^k$ is a vector in the normal cone to the face of F containing the observed sufficient statistics in its relative interior and, therefore, $\mathrm{supp}(\mathrm{B}\mathbf{z}^k)$ is the co-facial set corresponding to a face containing F, possibly to F itself.

If, on the other hand, $\mathbf{y}^k = 0$, then $\mathrm{supp}(\mathrm{B}\mathbf{z}^{k-1})$ is maximal and, therefore, identifies the co-facial set corresponding to F (that is, $-\mathbf{z}^{k-1}$ must also be a point in the relative interior of the normal cone to F).

Finally, if $\mathrm{supp}(\mathbf{y}^k)^c = \varnothing$, then $\mathcal{I}_{\mathrm{B}}$ itself is found to be the co-facial corresponding to F. ■

Remark.
An equivalent but much less efficient way to computing the facial set based is to solve I linear programs, one for each row of the design matrix A:

$$\begin{aligned} & \max \mathbf{a}_i^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^\top \mathbf{t} = 0 \\ & \mathrm{A}\mathbf{x} \geqslant \mathbf{0} \\ & -\mathbf{1} \leqslant \mathbf{x} \leqslant \mathbf{1} \end{aligned}$$

where $\mathbf{a}_i$ denotes the i-th row of A. Letting $\mathbf{x}_i$ denote the optimal solution to the i-th program, the MLE does not exist if and only if $\mathbf{a}_i^\top \mathbf{x}_i > 0$ for some i, in which case the facial set associated with $\mathbf{t}$ is given by

$$\{i \colon \mathbf{a}_i^\top \mathbf{x}_i = 0\}.$$

Geyer (2009) discusses a similar algorithm.

2.2 Newton-Raphson Procedure

We now describe a non-linear optimization approach to solve (5) using the Newton-Raphson method. While this procedure is also guaranteed to correctly return the appropriate facial set, it needs to be run only once, unlike the LP method presented above.

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

$$f(\mathbf{x}) = -\mathbf{1}^\top \exp(\mathrm{B}\mathbf{x}), \tag{8}$$

with gradient $\nabla f(\mathbf{x}) = -\mathrm{B}^\top \exp(\mathrm{B}\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x}) = -\mathrm{B}^\top \exp(\mathrm{B}\mathbf{x})\mathrm{B}$. The following proposition relates the problem of optimizing f with the existence of the MLE. In addition, when the MLE is nonexistent, the sequence of points $\{\mathbf{x}_n\}$ realizing the supremum of f is not only diverging, but it is guaranteed to eventually identify the appropriate co-face.

Proposition 2.6. *Let f be as in (8) and consider the optimization problem*

$$\sup_{\mathbf{x} \in \mathbb{R}^k} f(\mathbf{x}). \tag{9}$$

The MLE exists if and only if the maximum of the problem (9) is attained for a finite vector $\mathbf{x}^* \in \mathbb{R}^k$. If the MLE does not exist, for any optimizing sequence $\{\mathbf{x}_n^*\}$ such that $\sup_{\mathbf{x} \in \mathbb{R}^k} f(\mathbf{x}) = \lim_n f(\mathbf{x}_n^*)$,

$$\mathcal{I}_\mathrm{B} \backslash \mathrm{supp}(\lim_n \exp(\mathrm{B}\mathbf{x}_n^*)) = \mathcal{F}^c.$$

Proof. The function $f(\mathbf{x})$ is bounded from above and, since the Hessian is negative definite for each $\mathbf{x} \in \mathbb{R}^q$ (due to the fact that $\mathrm{rank}(\mathrm{B}) = q$), concave on $\mathbb{R}^q$. Thus, the optimum is unique and furthermore, it is enough to consider the first order optimality conditions. Suppose the optimum occurs for some vector $\mathbf{x}^* \in \mathbb{R}^k$, so that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Letting $\mathbf{y}^* = \exp(\mathrm{B}\mathbf{x}^*) > \mathbf{0}$, the optimality condition on the gradient implies that $\mathrm{B}^\top \mathbf{y}^* = \mathbf{0}$. By Stiemke's theorem 6.3, the system $\mathrm{B}\mathbf{x} \gneqq \mathbf{0}$ has no solutions hence the MLE exists. To show the reverse, assume the MLE exists, so the system $\mathrm{B}\mathbf{x} \gneqq \mathbf{0}$ is unfeasible. Then, by Stiemke's theorem again, the system $\mathrm{B}^\top \mathbf{y} = \mathbf{0}$ does not admit any positive solutions, which implies that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, for all $\mathbf{x} \in \mathbb{R}^q$.

To prove the second claim, suppose the MLE does not exists. Denote with $\mathbf{b}_i$ the i-th row of B. Then, there exists a subset (possibly improper) $\mathcal{F}^c$ of the row indices $\mathcal{I}_\mathrm{B}$ and a sequence $\{\mathbf{w}\}_n$ such that $\mathbf{b}_i^\top \mathbf{w}_n < 0$ for each n and $\mathbf{b}_i^\top \mathbf{w}_n \downarrow -\infty$ if $i \in \mathcal{F}^c$, while $\mathbf{b}_i^\top \mathbf{w}_n = 0$ for each n if $i \notin \mathcal{F}^c$. By concavity of f, there exists an optimizing sequence $\{\mathbf{x}_n^*\} \subset \mathbb{R}^q$ such that

$$\lim_n f(\mathbf{x}_n) = \sup_{\mathbf{x} \in \mathbb{R}^q} f(\mathbf{x}), \tag{10}$$

where $\lim_n \|\mathbf{x}_n^*\| = \infty$. Let $\mathbf{y}^* = \lim_n \exp(\mathrm{B}\mathbf{x}_n^*)$. It is easy to see that $\mathbf{y}^*$ does not depend on the choice of the optimizing sequence. Hence it is unique. We show that $\mathrm{supp}(\mathbf{y}^*)^c = \mathcal{I}_{\mathrm{B}} \backslash \mathrm{supp}(\mathbf{y}^*) = \mathcal{F}^c$, which will prove the claim. For any $i \in \mathrm{supp}(\mathbf{y}^*)^c$, it must be the case that $\mathbf{b}_i^\top \mathbf{x}_n^* < 0$ for all n big enough. This implies that $i \in \mathcal{F}^c$. Thus, we have established that $\mathrm{supp}(\mathbf{y}^*)^c \subseteq \mathcal{F}^c$. To show the opposite inclusion $\mathcal{F}^c \subseteq \mathrm{supp}(\mathbf{y}^*)^c$, suppose there exists an index $i \in \mathcal{F}^c$ which does not belong to $\mathrm{supp}(\mathbf{y}^*)^c$. Then, letting $\{\mathbf{w}\}_n$ be defined as above, $\mathbf{b}_i^\top \mathbf{w}_n \downarrow -\infty$ but $\lim_n |\mathbf{b}_i^\top \mathbf{x}_n^*| < \infty$, so that

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

which contradicts (10). Thus, $\mathcal{F}^c \subseteq \mathrm{supp}(\mathbf{y}^*)^c$. ■

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

As we already mentioned, we can optimaze the function (8) using Newton-Raphson method. In fact, when the MLE exists, (8) satisfies the conditions guaranteeing quadratic convergence (see Boyd and Vandenberghe, 2004, Chapter 9). When the MLE does not exists, we can still apply the Newton-Raphson procedure and it will return the correct facial set by producing a divergent Newton sequence, as we show in our next result.

Theorem 2.7. *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a strictly concave function of class $\mathcal{C}^3$, strongly concave on any bounded ball and having no maximum on the closure of the open ball B. For any $\mathbf{x} \in B$ let $\mathbf{d}_{\mathbf{x}}$ be the Newton direction corresponding to $\mathbf{x}$. Then, there exists a positive number $\alpha \leqslant 1$ such that such that*

$$f(\mathbf{x} + \alpha \mathbf{d}_{\mathbf{x}}) - f(\mathbf{x}) \geqslant \gamma, \tag{11}$$

for all $x \in B$, where $\gamma = \tau \left(\inf_{\mathbf{x} \in B} ||\nabla f(\mathbf{x})||^2\right)$ for some number τ depending on α and B only.

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

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

Using strict concavity on B' of f, there exist positive constants K and L such that

$$K \leqslant -\mathbf{y}^\top \nabla^2 f(\mathbf{x})^{-1} \mathbf{y} \leqslant L \tag{12}$$

for all $\mathbf{x} \in \bar{B}$ and all unit vectors $\mathbf{y}$. Since, for any $x \in B$,

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

it follows that

$$K||\nabla f(\mathbf{x})|| \leqslant ||\mathbf{d}_{\mathbf{x}}|| \leqslant L||\nabla f(\mathbf{x})||. \tag{13}$$

Let $\beta \in (0, 1]$, to be chosen below. Using Taylor's expansion,

$$f(\mathbf{x} + \beta \mathbf{d}_{\mathbf{x}}) = f(\mathbf{x}) + \beta \nabla f(\mathbf{x})^\top \mathbf{d}_{\mathbf{x}} + \frac{\beta^2}{2} \mathbf{d}_{\mathbf{x}}^\top \nabla^2 f(\mathbf{x} + c\beta \mathbf{d}_{\mathbf{x}}) \mathbf{d}_{\mathbf{x}}, \tag{14}$$

for some $0 < c < 1$. Using (12) and (13), we can bound the right hand side of (14). In fact,

$$\begin{array}{rcl} \beta \nabla f(\mathbf{x})^\top \mathbf{d}_{\mathbf{x}} & = & -\beta \frac{\nabla f(\mathbf{x})^\top}{\|\nabla f(\mathbf{x})\|} \nabla^2 f(\mathbf{x}) \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} \|\nabla f(\mathbf{x})\|^2 \\ & \geqslant & \beta K \|\nabla f(\mathbf{x})\|^2 \end{array}$$

and

$$\begin{array}{rcl} \frac{\beta^2}{2} \mathbf{d}_{\mathbf{x}}^\top \nabla^2 f(\mathbf{x} + c\beta \mathbf{d}_{\mathbf{x}}) \mathbf{d}_{\mathbf{x}} & = & \frac{\beta^2}{2} \frac{\mathbf{d}_{\mathbf{x}}^\top}{\|\mathbf{d}_{\mathbf{x}}\|} \nabla^2 f(\mathbf{x} + c\beta \mathbf{d}_{\mathbf{x}}) \frac{\mathbf{d}_{\mathbf{x}}}{\|\mathbf{d}_{\mathbf{x}}\|} \|\mathbf{d}_{\mathbf{x}}\|^2 \\ & \geqslant & -\frac{\beta^2}{2} L \|\mathbf{d}_{\mathbf{x}}\|^2 \\ & \geqslant & -\frac{\beta^2}{2} L^2 \|\nabla f(\mathbf{x})\|^2. \end{array}$$

Therefore,

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

The term $\left(\beta K - \frac{\beta^2}{2}L^2\right)$ is positive provided $\beta < \frac{2K}{L}$. Choose any $0 < \beta < \frac{2K}{L}$ and set $\alpha = \min\{1,\beta\}$ and $\tau = \left(\alpha K - \frac{\alpha^2}{2}L^2\right)$. Next, the term $\gamma = \tau\left(\inf_{\mathbf{x}\in B}||\nabla f(\mathbf{x})||^2\right)$ is strictly positive, as $\tau > 0$, f has no maximum on the closure of B. Therefore, for such a choice of α and γ, $f(\mathbf{x}+\alpha\mathbf{d}_{\mathbf{x}}) - f(\mathbf{x}) \geqslant \gamma$, as desired. ■

3 Maximization of the Extended Log-Likelihood Function

We describe the use of the Newton-Raphson algorithm to optimize of the log-likelihood function under natural parametrization (see also Haberman, 1974, Chapter 3). Provided the MLE exists, it is well known (see, for example, Agresti, 2002) that Newton-Raphson method for maximum likelihood estimation of the natural parameters eventually achieves a quadratic rate of convergence to its unique optimizer. However, when MLE fails to exist, the procedure becomes highly unstable. Indeed, in this case the negative log-likelihood function has directions of recessions and its optimum is realized as the limit of sequences with norms exploding to infinity. Furthermore, the Fisher information matrix evaluated along any optimizing sequence will converge to a singular matrix (as shown the remarks following corollary 8 of RF). However, as we explained above, these issues disappear once we optimize the extended log-likelihood function, as described below.

For hierarchical log-linear models and, in fact, for more general log-linear models, iterative algorithms such as the iterative proportional fitting or scaling (IPS), are popular, e.g., see Fienberg (1970), Bishop et al. (1975), Darroch and Ratcliff (1972), Csiszár (1975, 1989), Lauritzen (1996) and Ruschendorf (1995). See also Hunter (2004) for the variation known as MM algorithms. Typically, these alternative algorithms are very simple to implement and do not require matrix inversion, resulting in a much lower space complexity compared with the Newton-Raphson procedure. Furthermore, since they carry out optimization in the mean value space, convergence to the unique optimum occurs regardless of the existence of the MLE. These algorithms suffer from serious drawbacks, however, that do not affect the Newton-Raphson procedure. First, the rate of convergence is often unknown and can be extremely slow, especially if the MLE fails to exist. Secondly, it is typically hard to detect whether the MLE exists or not, other than by monitoring the (usually slow) rate to convergence. As a result, computing the number of estimable parameters (i.e. the dimension of the extended model) is rather difficult. Finally, these algorithms are suitable for estimating the mean-value parameters parameters, but not the natural parameters.

We cannot recommend one type of algorithm over the other for all practical purposes: this choice necessarily depends on a various factors, such as the dimension of the problem, the computational resources available and the time constraints, all of which contributing to the trade-off between space complexity (low for IPS and large for Netwon-Raphson) versus time complexity (high for IPS and low for Netwon-Raphson). For very high-dimensional problems, however, IPS algorithms may be the only feasible way.

A notable instance where IPS confers computation simplifications is for decomposable log-linear models, for which the MLE and extended MLE are rational functions of the table margins (see, e.g., Haberman, 1974; Lauritzen, 1996; Geiger et al., 2006). For these models, the IPS algorithm is known to converge in very few iterations; in fact, provided that the IPS updates are carried out according to a perfect ordering of the graph cliques, convergence is achieved in just one pass. Furthermore, explicit formulae for computing the number of estimable parameters even with a nonexistent MLE are available (see Lauritzen, 1996). Thus, for decomposable log-linear models, we can carry out extended maximum likelihood estimation very efficiently.

For non-decomposable models, there are a number of modifications of the IPS algorithm that guarantee in some cases a considerable reduction in computational complexity. Most of these improvements rely on graph theoretical properties of the log-linear models and on some form of approximation of the original model by decomposable models. See, in particular, Jiroušek and Přeučil (1995); Jiroušek (1991), Badsberg and Malvestuto (2001) and Endo and Takemura (2009).

We can optimize the Newton-Rapshon algorithm for hierarchical log-linear models, through a careful handling of the design matrices. Indeed, we can construct these matrices to be sparse or have rather special structures, properties that can be exploited to significantly reduce the computational burden associated with matrix multiplication and inversion (see Fienberg et al., 1980).

3.1 Poisson Sampling Scheme

Using a full-rank design matrix A, the MLE of the natural parameter and the mean-value parameters for the corresponding log-linear model can be found by optimizing the function

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \ell^P(\mathbf{x}),$$

where $\ell^P(\mathbf{x}) = \mathbf{n}^\top \mathrm{A}\mathbf{x} - \mathbf{1}^\top \exp(\mathrm{A}\mathbf{x})$. If the MLE exists, the optimum will be attained by a unique vector $\widehat{\boldsymbol{\theta}}$ of finite norm, the MLE of the natural parameter. The MLE of the mean value parameter is then $\widehat{\mathbf{m}} = \exp(\mathrm{A}\widehat{\boldsymbol{\theta}}) > \mathbf{0}$.

The gradient and Hessian of ℓ^P, needed by Newton-Raphson algorithm, are easy to evaluate. Indeed, setting $\boldsymbol{\mu}_\mathbf{x} = \mathrm{A}\mathbf{x}$ and $\mathbf{m}_\mathbf{x} = \exp(\boldsymbol{\mu}_\mathbf{x})$, it can be seen that

$$\begin{aligned} \ell^P(\mathbf{x}) &= \mathbf{n}^\top \boldsymbol{\mu}_\mathbf{x} - \mathbf{1}^\top \mathbf{m}_\mathbf{x} \\ \nabla \ell^P(\mathbf{x}) &= \mathrm{A}^\top(\mathbf{n} - \mathbf{m}_\mathbf{x}) \\ \nabla^2 \ell^P(\mathbf{x}) &= -\mathrm{A}^\top \mathrm{D}_{\mathbf{m}_\mathbf{x}} \mathrm{A}, \end{aligned} \tag{15}$$

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

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

At the k-th step of the Newton Raphson algorithm, we can use the current approximation $\mathbf{x}_k$, along with Equation (15), to compute the Newton direction $\mathbf{d}_k$ by solving the system

$$\nabla^2 \ell^P(\mathbf{x}_k)\mathbf{d}_k = \nabla \ell^P(\mathbf{x}_k). \tag{16}$$

The Cholesky factorization of $\nabla^2 \ell^P(\mathbf{x}_k)$ is useful for solving the above system.

After we compute the new direction by solving the system (16), we must determine the stepsize α_k. To this end, we consider the scalar function

$$\phi_k(\alpha) = \ell^P(\mathbf{x}_k + \alpha \mathbf{d}_k),$$

and set $\mathbf{c}_k = \mathrm{A}\mathbf{d}_k$, so that

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

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

$$\phi_k'(\alpha) = \mathbf{c}_k^\top \left(\mathbf{n} - (\mathbf{m}_k \cdot \exp(\alpha \mathbf{c}_k))\right) \quad \text{and} \quad \phi_k''(\alpha) = -\sum_i \mathbf{c}_k^2(i)\mathbf{m}_k(i)\exp(\alpha \mathbf{c}_k(i)).$$

With this information, we can compute the step-size α_k using any of the usual strategies (see, for instance Boyd and Vandenberghe, 2004, Chapter 9).

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

When the MLE does not exist because $\mathbf{t}$ belongs to the relative interior of a face F of C_{A} of dimension d_F, and the extended MLE corresponds to a facial set $\mathcal{F}$, the extended log-likelihood optimization problem becomes

$$\sup_{\mathbf{w} \in \mathbb{R}^{d_F}} \ell^P_{\mathcal{F}}(\mathbf{x}), \tag{17}$$

where $\ell^P_{\mathcal{F}}(\mathbf{x}) = \mathbf{n}^\top \mathrm{A}^*_{\mathcal{F}} \mathbf{x} - \mathbf{1}^\top \exp(\mathrm{A}^*_{\mathcal{F}} \mathbf{x})$ and $\mathrm{A}^*_{\mathcal{F}}$ is a $|\mathcal{F}| \times d_F$ full-column rank design matrix consisting of any set of linearly independent columns from $\mathrm{A}_{\mathcal{F}}$. In order to compute $\mathrm{A}^*_{\mathcal{F}}$ from $\mathrm{A}_{\mathcal{F}}$ proposition 5.1 in section 5 could be used, for instance. The optimum $\mathbf{x}^*$ for (17) is the extended MLE of the natural parameters, while the extended MLE for the cell mean values is the non-negative vector

$$\widehat{\mathbf{m}}^{\mathrm{e}} = \tau_{\mathcal{F}}\left(\exp(\mathrm{A}^*_{\mathcal{F}} \mathbf{x}^*)\right).$$

3.2 Product Multinomial Sampling Scheme

We now provide the details for carrying out maximum likelihood estimation under product multinomial setting. When the product multinomial sampling scheme applies, two strategies for maximum likelihood estimation are available. One possibility is to take advantage that the extended MLE of the cell mean values are identical under Poisson and product multinomial scheme, and, provided the sampling subspace $\mathcal{N}$ is contained in the log-linear subspace $\mathcal{M}$, proceed as if Poisson sampling were in fact used. On the one hand, this approach is appealing because, as we just saw, the computations for the Poisson log=likelihood are relatively straightforward and computationally inexpensive; on the other hand, those computations are carried out over d-dimensional space, while the effective number of parameters is $d - m = \dim(\mathcal{M} \ominus \mathcal{N})$. The second possibility is to use lemma 2 of RF and thus optimize the log-likelihood function parametrized in minimal form by any full-rank design matrix for $\mathcal{M} \ominus \mathcal{N}$. This second approach is more elaborated because, as we will see below, the gradient and Hessian of the re-parametrized log-likelihood are more complicated and harder to evaluate numerically. When $\dim(\mathcal{N})$ is very large relative to to d, however, we can achieve a considerable reduction in the dimensionality, more than offsetting the computational ease of the Poisson case, despite the increase in complexity needed to obtain the Newton steps.

Before we proceed, we show how to obtain a design matrix for $\mathcal{M}$. Let A_1 be the $I \times m$ matrix whose j-column is $\boldsymbol{\chi}_j$ (see section 2 of RF), so that $\mathrm{D} = \mathrm{A}_1^\top \mathrm{A}_1$ is a m-dimensional non-singular diagonal matrix. Let $\mathrm{A} = [\mathrm{A}_1\ \mathrm{A}_2]$ be such that $\mathcal{R}(\mathrm{A}) = \mathcal{M}$.

Lemma 3.1. *The columns of the matrix* $\mathrm{W} = \mathrm{A}_2 - \mathrm{A}_1 \mathrm{D}^{-1} \mathrm{A}_1^\top \mathrm{A}_2$ *form a basis for* $\mathcal{M} \ominus \mathcal{N}$.

Proof. Orthogonality of $\mathcal{R}(\mathrm{W})$ and $\mathcal{R}(\mathrm{A}_1)$ follows from the chain of equalities

$$\begin{aligned} \mathrm{A}_1^\top \mathrm{W} &= \mathrm{A}_1^\top \mathrm{A}_2 - \mathrm{A}_1^\top \mathrm{A}_1 \mathrm{D}^{-1} \mathrm{A}_1^\top \mathrm{A}_2 \\ &= \mathrm{A}_1^\top \mathrm{A}_2 - \mathrm{D}\mathrm{D}^{-1} \mathrm{A}_1^\top \mathrm{A}_2 \\ &= 0. \end{aligned} \tag{18}$$

It only remains to show that $(\mathrm{A}_1\ \mathrm{W})$ span $\mathcal{M}$. Let $\boldsymbol{\mu} = \mathrm{A}_1 \mathbf{b}_1 + \mathrm{A}_2 \mathbf{b}_2$ be any vector in $\mathcal{M}$. Then,

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

so that $\boldsymbol{\mu}$ is a linear combination of the columns of A_1 and of W. ■

Determination of the facial set

As we described in section 2, we can determine the facial set by solving (5). Under a product multinomial scheme,we modify the procedure to obtain matrix B as follows. Assume that the design matrix for $\mathcal{M}$ has the form $\mathrm{A} = [\mathrm{A}_1 \;\; \mathrm{A}_2]$ described above and set

$$\mathrm{W} = \mathrm{A}_2 - \mathrm{A}_1 \mathrm{D}_+^{-1} \mathrm{A}_{1,+}^\top \mathrm{A}_{2,+},$$

where $\mathrm{A}_{1,+}$ and $\mathrm{A}_{2,+}$ are the submatrices of A_1 and A_2 obtained considering only the rows indexed by $\mathcal{I}_+$, respectively, and $\mathrm{D}_+ = \mathrm{A}_{1,+}^\top \mathrm{A}_{1,+} = \mathrm{D}$, diagonal and invertible. Using the very same arguments from the proof of lemma 3.1, we see that the columns of $[\mathrm{A}_1 \; \mathrm{W}]$ span $\mathcal{M}$ and that the columns of W_+ are orthogonal to the columns of $\mathrm{A}_{1,+}$. It follows from the independence of the columns of $\mathrm{A}_{1,+}$ that any basis for the null space of $[\mathrm{A}_1 \; \mathrm{W}]_+$ must have the form

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

i.e. the entire dependency resides in the columns of V_+. We can now use the matrix

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

to set up the optimization problem (5) for the determination of facial sets, after the elimination of possible redundant zero rows.

Optimization of the extended log-likelihood

According to lemma 2 of RF, we can parametrize the log-likelihood using the vectors $\boldsymbol{\beta} \in \mathcal{M} \ominus \mathcal{N}$. In fact, with a slight abuse of notation, we can write the log-likelihood function as

$$\ell^M(\mathbf{x}) = \mathbf{n}^\top \mathrm{W}\mathbf{x} - \sum_{j=1}^{m} N_j \log \boldsymbol{\chi}_j^\top \exp(\mathrm{W}\mathbf{x}), \tag{19}$$

where, in Poisson sampling case, $\ell^M(\mathbf{x})$ is in fact identical to $\ell^M(\boldsymbol{\beta})$ from equation 9 of RF, with $\boldsymbol{\beta} = \mathrm{W}\mathbf{x}$.

For $\mathbf{x} \in \mathbb{R}^{d-m}$, let $\mathbf{b}_\mathbf{x} = \exp(\boldsymbol{\beta}_\mathbf{x})$, where $\boldsymbol{\beta}_\mathbf{x} = \mathrm{W}\mathbf{x}$ (in fact, W is a homeomorphism between $\mathcal{M} \ominus \mathcal{N}$ and $\mathbb{R}^{d-m}$). The proof of lemma 2 of RF shows that, for each $\boldsymbol{\beta}_\mathbf{x}$ there exists a corresponding $\boldsymbol{\nu}_\mathbf{x} \in \mathcal{N}$ such that the vector $\mathbf{c}_\mathbf{x} = \exp(\boldsymbol{\nu}_\mathbf{x})$ satisfies

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

Then, some algebra yields that the gradient at $\mathbf{x}$ is

$$\nabla \ell^M(\mathbf{x}) = \mathrm{W}^\top \mathbf{n} - \mathrm{W}^\top \begin{pmatrix} \left(\frac{N_1}{\boldsymbol{\chi}_1^\top \mathbf{b}_\mathbf{x}}\right) \mathbf{b}_\mathbf{x}^1 \\ \vdots \\ \left(\frac{N_r}{\boldsymbol{\chi}_r^\top \mathbf{b}_\mathbf{x}}\right) \mathbf{b}_\mathbf{x}^m \end{pmatrix} = \mathrm{W}^\top \mathbf{n} - \mathrm{W}^\top \mathbf{m}_\mathbf{x}.$$

while the Hessian is

$$\begin{array}{rcl} \nabla^2 \ell^M(\mathbf{x}) & = & -\mathrm{D}_{\mathbf{m}_\mathbf{x}} - \sum_{j=1}^m \frac{1}{N_j} \mathbf{m}_\mathbf{x}^j (\mathbf{m}_\mathbf{x}^j)^\top \\ & = & -\mathrm{D}_{\mathbf{m}_\mathbf{x}} \left(\mathrm{I} - \Pi_\mathcal{N}^{\mathbf{m}_\mathbf{x}}\right), \end{array}$$

where $\mathbf{m}_\mathbf{x}^j = \{\mathbf{m}_\mathbf{x}(i) \colon i \in \boldsymbol{\chi}_j\}$ for $j = 1, \ldots, m$ and $\Pi_\mathcal{N}^\mathbf{m}$ is the (oblique) orthogonal projection matrix onto $\mathcal{N}$ relative to the inner product $[\cdot,\cdot]_\mathbf{m}$ on $\mathbb{R}^\mathcal{I}$ defined by $[\mathbf{x}, \mathbf{y}] = \mathbf{x}^\top \mathrm{D}_\mathbf{m} \mathbf{y}$ (see equation 2.28 in Haberman, 1974). For a characterization of maximum likelihood estimation in terms of oblique projections on $\mathcal{M} \ominus \mathcal{N}$, see Haberman (1977). Note that $\nabla^2 \ell^M$ is negative definite on $\mathbb{R}^{d-m}$ but not strongly concave on it. As in the Poisson case, this feature allows for the possibility of a non-existent MLE. An equivalent expression for the Hessian is

$$\nabla^2 \ell^M(\mathbf{x}) = -\sum_{j=1}^{m} \mathrm{W}_j^\top \mathrm{H}_\mathbf{x}^j \mathrm{W}_j,$$

where W_j denotes the submatrix of W obtained by considering only the rows indexed by $\mathrm{supp}(\boldsymbol{\chi}_j)$ and

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

When the MLE does not exist and the extended MLE corresponding to for a given facial set $\mathcal{F}$, the procedure is identical to the Poisson case. Specifically, we re-define the restricted log-likelihood function (19) with domain $\mathbb{R}^{d_F}$, $d_F < d - m$, as

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

where $\mathrm{W}^*_\mathcal{F}$ is the full-column-rank $|\mathcal{F}| \times d_F$ dimensional matrix consisting of any set of linearly independent columns of $\mathrm{W}_\mathcal{F}$, isolated using any of the procedures described in section 5. As usual, d_F is the dimension of the face of the convex support of the associated family containing the observed sufficient statistics $\mathrm{A}^\top \mathbf{n}$ in its relative interior (see theorem 3 of RF). The extended MLE of the natural parameters is the unique solution $\mathbf{x}^*$ to the optimization problem

$$\sup_{\mathbf{x} \in \mathbb{R}^{d_F}} \ell^M_\mathcal{F}(\mathbf{x}).$$

The extended MLE for the cell mean vectors is then vector $\hat{\mathbf{m}}^\mathrm{e} \in \mathbb{R}^{\mathcal{I}_{\geqslant 0}}$ with coordinates

$$\hat{\mathbf{m}}^\mathrm{e}(i) = \begin{cases} \mathbf{b}_{\mathbf{x}^*}(i) \mathbf{c}_{\mathbf{x}^*}(i) & \text{if } i \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathbf{b}_{\mathbf{c}^*} = \exp(\mathrm{W}^*_\mathcal{F} \mathbf{x}^*)$ and $\mathbf{c}_{\mathbf{x}^*}$ is the vector $\mathbb{R}^\mathcal{I}$ with coordinates $\mathbf{c}_{\mathbf{x}^*}(i) = \frac{N_j}{\boldsymbol{\chi}_j^\top \mathbf{b}_{\mathbf{x}^*}}$, for $i \in \boldsymbol{\chi}_j$, $j = 1 \ldots, m$.

4 The case $\mathcal{N} \not\subset \mathcal{M}$

We have followed the convention, suggested by Haberman (1974), that $\mathcal{N} \subsetneq \mathcal{M}$. In practice, there are certainly log-linear models for which this assumption fails, such as the Rasch model or variations on the Bradley-Terry paired comparisons model. For simplicity we consider only the case that $\mathcal{N}$ is the sampling subspace generated by a product multinomial sampling scheme and that $\mathcal{N} \cap \mathcal{M} = \{\mathbf{0}\}$.

In order to check for the existence of the MLE and to determine the appropriate facial sets, we can use the following trick, which consists in building a larger model for which the calculations are simpler. Set $\mathcal{M}' = \mathcal{M} + \mathcal{N} \subset \mathbb{R}^\mathcal{I}$ and consider a new, $(d+m)$-dimensional log-linear model with log-linear subspace $\mathcal{M}'$ and product multinomial sampling scheme with sampling subspace $\mathcal{N}$. Then, if A is a design matrix for $\mathcal{M}$ of dimension $\mathcal{I} \times d$ and V the sampling matrix for $\mathcal{N}$ of dimension $\mathcal{I} \times m$, $\mathrm{A}' = (\mathrm{A},\ \mathrm{V})$ and $C_{\mathrm{A}'}$ are the design matrix for $\mathcal{M}'$ and its marginal cone, respectively. Thus, if $\mathbf{t} \in \mathbb{R}^d$ is the observed sufficient statistics for $\mathcal{M}$, $\mathbf{t}' = (\mathbf{t}, \mathbf{1})$ is also sufficient. Then, by theorem 3 of RF, the MLE for the parameters of $\mathcal{M}$ exists if and only if $\mathbf{t}' \in \mathrm{ri}(C_{\mathrm{A}'})$ and, furthermore, the facial set associated with $\mathbf{t}$ is precisely the facial set of $C_{\mathrm{A}'}$ associated with $\mathbf{t}'$. For these tasks, we can use the algorithms of section 2.

The reason why it is simpler to work with the enlarged model $\mathcal{M}'$ is that we do all the calculations on the corresponding marginal cone, and not on the convex support for the original log-linear model $\mathcal{M}$. Indeed, under product multinomial scheme, this convex support takes the form of a Minkwoski sum of polytopes, which is computationally very hard to manage. In contrast, the marginal cone C'_A, though having larger dimension, is much simpler to handle both theoretically and computationally. In polyhedral geometry, this trick goes under the name of Cayley embedding. See Rinaldo et al. (2011) for a detailed application of this trick to the problem of existence of the MLE for generalized Bradley-Terry and Rasch models and for the determination of the relevant facial sets.

For the maximization of the extended likelihood, there are two options. The first is to optimize directly the extended product multinomial log-likelihood for the model $\mathcal{M}$ (see section 3.2 for details), and the other option is to optimize the extended Poisson log-likelihood of the enlarged model $\mathcal{M}'$. By theorem 7 of RF, the resulting estimates coincide. The advantage of the first method is that is has a smaller number of parameters to optimize, while the advantage of the second method is that the likelihood itself, though depending on a larger set of parameters, is simpler to optimize.

5 Detecting Rank Degeneracies

The present section describes a method for isolating a set of independent columns from a matrix U based on the Cholesky decomposition with pivoting. See Stewart (1998) for detailed descriptions and properties of algorithms we use below.

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

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

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

$$\mathrm{U} - \mathrm{R}_k^\top \mathrm{R}_k = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{U}_k \end{pmatrix}, \tag{20}$$

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

A simple modification of the algorithm described above allows us to consider matrices that are only positive semidefinite. In fact, it is not necessary to accept diagonal elements as pivots (i.e. as determining the diagonal elements of R). Specifically, suppose that, at the k-th stage of the reduction algorithm represented by equation (20), the pivoting for the next stage is obtained using another diagonal entry of U_k, say $a_{l,l}^{(k)}$, $l \neq 1$, instead of $a_{1,1}^{(k)}$. Let $\mathrm{J}'_{k+1,l}$ be a permutation matrix obtained by exchanging the first and l-th rows of the identity matrix of order $p-k$ so that

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

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

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

Then, from (20),

$$\mathrm{J}_k\mathrm{U}\mathrm{J}_k - \mathrm{J}_k\mathrm{R}_k^\top\mathrm{R}_k\mathrm{J}_k = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{J}'_{k+1,l}\,\mathrm{U}_k\,\mathrm{J}'_{k+1,l} \end{pmatrix}. \tag{21}$$

The matrix $\mathrm{R}_k\mathrm{J}_k$ differs from R_k only in having its $(k+1)$-th and $(k+l)$-th columns interchanged. Consequently, (21) represent the k-th step of the Cholesky decomposition of $\mathrm{J}_k\mathrm{U}\mathrm{J}_k$ in which $a_{1,1}^{(k)}$ has been replaced by $a_{l,l}^{(k)}$. If interchanges of leading terms are made at each step, with the exception of the last one, the Cholesky factorization will produce an upper triangular matrix R such that

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

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

If U is positive semidefinite and we carry the algorithm to its k-th stage, we can show that U_k is also positive semidefinite. Unless U_k is zero, it will have a positive diagonal element, which we can exchange into the pivot to initiate the $(k+1)$ step. Among the possible pivoting strategies, one that is particularly well-suited to problems of rank detection is taking as pivot element the largest diagonal element of U_k, for every stage k of the reduction. This will result in a matrix R such that

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

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

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

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

The following result show how we can the Cholesky decomposition with pivoting of a positive semidefinite matrix to isolate a set of independent columns from a matrix A

Proposition 5.1. *Let* A *be a matrix with* p *columns and* J *be a permutation matrix such that* $\mathrm{AJ} = [\mathrm{A}_1\ \mathrm{A}_2]$ *with* A_1 *having* k *columns. If*

$$(\mathrm{A}_1, \mathrm{A}_2)^\top (\mathrm{A}_1, \mathrm{A}_2) = \begin{pmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} \\ 0 & 0 \end{pmatrix}^\top \begin{pmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} \\ 0 & 0 \end{pmatrix},$$

where R_{11} *is non-singular of order* k*, then:*

i. the columns of A_1 *are linearly independent;*

ii. $\mathrm{A}_2 = \mathrm{A}_1 \mathrm{R}_{11}^{-1} \mathrm{R}_{12}$*;*

iii. the columns of the matrix

$$\begin{pmatrix} -\mathrm{R}_{11}^{-1}\mathrm{R}_{12} \\ \mathrm{I}_{p-k} \end{pmatrix} \tag{22}$$

form a basis for the null space of UJ.

Proof. Since $\mathrm{A}_1^\top \mathrm{A}_1 = \mathrm{R}_{11}^\top \mathrm{R}_{11}$, and R_{11} is non-singular and positive definite, A_1 has independent columns, proving *i.*. To establish *ii.*, note that

$$\operatorname{rank}(\mathrm{A}_1\ \mathrm{U}_2) = \operatorname{rank} \begin{pmatrix} \mathrm{R}_{11} & \mathrm{A}_{12} \\ 0 & 0 \end{pmatrix} = k,$$

so we can obtain A_2 as a linear combination of columns of A_1:

$$\mathrm{A}_2 = \mathrm{A}_1 \mathrm{X}. \tag{23}$$

Then, after pre-multiplying both sides by $\mathrm{A}_1^\top$,we get

$$\mathrm{X} = (\mathrm{A}_1^\top \mathrm{A}_1)^{-1} \mathrm{A}_1^\top \mathrm{A}_2 = \mathrm{R}_{11}^{-1} \mathrm{R}_{12}.$$

To show *iii*, we observe that the matrix (22) has $p-k$ independent columns and, by *i*, it satisfies

$$(\mathrm{A}_1, \mathrm{A}_2) \begin{pmatrix} -\mathrm{R}_{11}^{-1}\mathrm{R}_{12} \\ \mathrm{I} \end{pmatrix} = -\mathrm{A}_1 \mathrm{R}_{12}^{-1} \mathrm{R}_{12} + \mathrm{A}_2 = 0,$$

where we justify the last inequality by (23). Since the null space of $(\mathrm{A}_1\ \mathrm{A}_2)$ has dimension $p-k$, the columns of (22) form a basis for it. ■

6 Alternative Methods for Determining Facial Sets

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

6.1 Maximum Entropy Approach

We can carry out identification of the appropriate facial set by replacing the linear objective function of the optimization problem (7), with Shannon's entropy function. The new problem is

$$\begin{array}{lccc} & \max -\sum_{i\in\mathcal{I}_{\mathrm{B}}}\mathbf{y}(i)\log\mathbf{y}(i) & & \\ \text{s.t.} & \mathbf{y} & = & \mathrm{B}\mathbf{x} \\ & \mathbf{y} & \geqslant & \mathbf{0} \\ & \mathbf{1}^\top\mathbf{y} & = & 1. \end{array} \tag{24}$$

The strictly concavity of the entropy function and the fact that $\lim_{x\downarrow 0} x\log x = 0$ guarantee that, for the unique maximizer $\mathbf{y}^*$ of 24, $\mathrm{supp}(\mathbf{y}^*)$ is maximal with respect to inclusion. In fact, letting Δ_0 denote the simplex in $\mathbb{R}^{\mathcal{I}_{\mathrm{B}}}$, we maximize the entropy function over the convex polytope $\mathrm{DC}_{\mathrm{B}}^{\mathbf{1}} := \mathrm{DC}_{\mathrm{B}} \cap \Delta_0$. Such intersection is trivial when the MLE exists and is the point $\mathbf{0}$. In this case, the problem is infeasible. Otherwise, due to the strict concavity of the entropy function, the optimum occurs inside $\mathrm{ri}\left(\mathrm{DC}_{\mathrm{B}}^{\mathbf{1}}\right)$, which corresponds to the maximal co-face. Note that $\mathrm{DC}_{\mathrm{B}}^{\mathbf{1}}$ is typically not of full dimension (unless $\mathcal{I}_{\mathrm{B}} = \mathcal{F}^c$), in which case the maximizer belongs to a relatively open neighborhood inside $\mathrm{DC}_{\mathrm{B}}^{\mathbf{1}}$.

If we denote the i-th row of B by $\mathbf{b}_i^\top$, we can rewrite the problem (24) in a more compact form by making the constraint $\mathbf{y} = \mathrm{B}\mathbf{x}$ implicit. Then

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

where, for $\mathrm{B}\mathbf{x} > \mathbf{0}$, $H(\mathbf{x}) = -\sum_{i\in\mathcal{I}_{\mathrm{B}}}\mathbf{b}_i^\top\mathbf{x}\log(\mathbf{b}_i^\top\mathbf{x})$, with gradient

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

and Hessian

$$\nabla^2 H(\mathbf{x}) = -\mathrm{B}^\top\mathrm{diag}\left(\mathrm{B}\mathbf{x}\right)^{-1}\mathrm{B} = -\sum_i \frac{1}{\mathbf{b}_i^\top\mathbf{x}}\mathbf{b}_i\mathbf{b}_i^\top.$$

6.2 Maximum Entropy and Newton-Raphson

By taking the log of the negative of the function f of Equation (8), we can represent the optimization problem (9) as an unconstrained geometric program

$$\min \quad \log\left(\sum_{i\in\mathcal{I}_{\mathrm{B}}}\exp(\mathbf{b}_i^\top\mathbf{x})\right),$$

which is equivalent to a linearly constrained one,

$$\begin{array}{ll} \min & \log\left(\sum_{i\in\mathcal{I}_{\mathrm{B}}}\exp(\mathbf{y}(i))\right) \\ \text{s.t.} & \mathrm{B}\mathbf{x} = \mathbf{y}, \end{array} \tag{25}$$

with feasible set given by the kernel of the matrix $(\mathrm{I}\ \ -\mathrm{B})$. See also Borwein and Lewis (2000, theorem 2.2.6). If $\mathbf{x}^*$ is the maximizer of the original problem (9), then this is also the minimizer for the geometric program (25), where the infimum can possibly be $-\infty$, but only when $\sup_{\mathbf{x}\in\mathbb{R}^k} f(\mathbf{x}) = 0$).

The conjugate of the log-sum-exp function appearing in (25) is the negative entropy function restricted to the simplex, given by

$$\begin{cases} \sum_i \boldsymbol{\nu}(i) \log \boldsymbol{\nu}(i) & \boldsymbol{\nu} \geqslant 0 \quad \mathbf{1}^\top \boldsymbol{\nu} = 1 \\ \infty & \text{otherwise,} \end{cases}$$

so the dual of the reformulated problem (25) is

$$\begin{array}{ll} \max & -\sum_{i \in \mathcal{I}_{\mathrm{B}}} \boldsymbol{\nu}(i) \log \boldsymbol{\nu}(i) \\ \text{s.t.} & \mathbf{1}^\top \boldsymbol{\nu} = 1 \\ & \mathrm{B}^\top \boldsymbol{\nu} = \mathbf{0} \\ & \boldsymbol{\nu} \geqslant \mathbf{0}. \end{array} \tag{26}$$

Proposition 6.1. *If the MLE exists, the problem (26) admits a unique strictly positive solution $\boldsymbol{\nu}^*$. If the MLE does not exist:*

a) if the zeros $\mathcal{I}_{\mathrm{B}} = \mathcal{F}^c$, then (26) is infeasible;

b) otherwise, the problem (26) is feasible and admits a unique solution $\boldsymbol{\nu}^*$ such that the co-face is given by the coordinates not in $\mathrm{supp}(\boldsymbol{\nu}^*)$.

Proof. Note that, by the properties of the entropy function, any solution $\boldsymbol{\nu}^*$ to the above problem has maximal support among all the non-negative vectors satisfying the equality constraint. Next, by strict concavity of the entropy function, if the problem is feasible, then it admits a unique solution. If the MLE exists, the maximum occurs at a strictly positive point $\boldsymbol{\nu}^* > \mathbf{0}$ by Stiemke's theorem 6.3.

Suppose instead that the MLE does not exist. If the system $\mathrm{B}\mathbf{x} > \mathbf{0}$ admits a solution, then, by Gordan's theorem 6.2, there is only one vector $\boldsymbol{\nu}^*$ satisfying the matrix equality constraint: $\boldsymbol{\nu}^* = \mathbf{0}$. Therefore, in this case the problem is infeasible. This proves *a)*. Otherwise, the solution is given by a vector $\boldsymbol{\nu}^* \gneqq \mathbf{0}$. In this case, the coordinates in $\mathrm{supp}(\boldsymbol{\nu}^*)^c$ give the appropriate co-face. In fact, $\mathbf{0} = (\boldsymbol{\nu}^*)^\top \mathrm{B} = (\boldsymbol{\nu}^*)^\top \mathrm{A}_0 \mathrm{X}$ implies that every $\mathbf{d} \in \mathrm{kernel}(\mathrm{A}_+)$ will be orthogonal to a strictly positive convex combination of the rows of A_0 corresponding to the coordinates in $\mathrm{supp}(\boldsymbol{\nu}^*)$. Since the MLE does not exist, there exists a vector $\mathbf{d}_* \in \mathrm{kernel}(\mathrm{A}_+)$ such that $\mathbf{d}_*^\top \mathbf{u}_i = 0$ for all $i \in \mathcal{F}$ and $\mathbf{d}_*^\top \mathbf{u}_i > 0$ for all $i \in \mathcal{F}^c$, that is, $\mathbf{d}_*$ is orthogonal to all strictly positive combinations of rows of A indexed by $\mathcal{F}$. By maximality of $\mathrm{supp}(\boldsymbol{\nu}^*)$, these rows are the ones in A_+ and the ones in $\mathrm{supp}(\boldsymbol{\nu}^*)$. Hence the result in *b)*. ■

Appendix: Theorems of Alternatives

Theorem 6.2 (Gordan's Theorem of Alternatives). *Given a matrix* A, *the following are alternatives:*

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

Theorem 6.3 (Stiemke's Theorem of Alternatives). *Given a matrix* A, *the following are alternatives:*

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

Schrijver (1998) provides proofs of both theorems.

References

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

Badsberg, J. and F. Malvestuto (2001). An implementation of the iterative proportional fitting procedure by propagation trees. *Computational Statistics and Data Analysis 37*, 297–322.

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

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press. Reprinted by Springer (2007).

Borwein, J. M. and A. S. Lewis (2000). *Convex Analysis and Nonlinear Optimization: Theory and Examples* (Second ed.). CMS Books in Mathematics. Springer.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization.* New York: Cambridge University Press.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability 3*(1), 146–158.

Csiszár, I. (1989). A geometric interpretation of darroch and ratcliff's generalized iterative scaling. *Annals of Statistics 17*(3), 1409–1413.

Darroch, J. N. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics 43*, 1470–1480.

Endo, Y. and A. Takemura (2009). Iterative proportional scaling via decomposable submodels for contingency tables. *Computational Statistics and Data Analysis 53*, 966–978.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics 41*(3), 907–917.

Fienberg, S. E., M. Meyer, and G. W. Stewart (1980). The numerical analysis of contingency tables. Unpublished manuscript.

Geiger, D., C. Meek, and B. Sturmfels (2006). On the toric algebra of graphical models. *Annals of Statistics 34*(3), 1463–1492.

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

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. *Annals of Statistics 5*(6), 1148–1169.

Hunter, D. (2004). Mm algorithms for generalized bradley-terry models. *Annals of Statistics 1*(32), 384–406.

Jiroušek, R. (1991). Solution of the marginal problem and decomposable distributions. *Kybernetika 27*, 403–412.

Jiroušek, R. and S. Přeučil (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics and Data Analysis 19*, 177–189.

Lauritzen, S. F. (1996). *Graphical Models.* New York: Oxford University Press.

Rinaldo, A., S. Fienberg, and Y. Zhou (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics 3*, 446–484.

Rinaldo, A., S. Petrović, and S. Fienberg (2011). Maximum likelihood estimation in network models. Technical report. available at http://arxiv.org/abs/1105.6145.

Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *Annals of Statistics 23*(4), 1160–1174.

Schrijver, A. (1998). *Theory of Linear and Integer Programming.* New York: Wiley & Sons.

Stewart, G. W. (1998). *Matrix Algorithms: Basic Decompositions*, Volume I. Society for Industrial and Applied Math.