Inference with sparse data in discrete hierarchical models: a Bayesian and a frequentist approach

Hélène Massam

June 20, 2016

Abstract

The material presented below is extracted from two papers:

- G. Letac and H. Massam. Bayes regularization and the geometry of discrete hierarchical loglinear models. Annals of Statistics, 40:861-890, 2012.
- N. Wang, J. Rauh and H. Massam. Approximating faces of marginal polytopes in discrete hierarchical models. Submitted for publication, (2016)

Keywords: existence of the maximum likelihood estimate, marginal polytope, faces, facial sets, extended maximum likelihood estimate.

1 Hierarchical models, discrete exponential families and their closure, polytopes

In the following four subsections, we recall some basic facts about hierarchical models, discrete exponential families, polytopes and the closure of exponential families, and we also define the extended mle.

1.1 Hierarchical models

For details and proofs on the material in this subsection, we refer to Letac& Massam(2012) and Rauh, Kahle & Ay. Let $X = (X_v, v \in V)$ be a discrete random vector with components indexed by $V = \{1, \ldots, p\}$, a finite set. Each variable X_v takes values in a finite set $I_v, v \in V$. The vector X takes its values in

$$I = \prod_{v \in V} I_v,$$

the set of cells $i = (i_v, v \in V)$ of the p-dimensional contingency table. Let Δ be a set of subsets of V which is a simplicial complex, that is, Δ is a set of subsets $D \subset V$ such that $D \in \Delta$ and $D' \subset D$ implies $D' \in \Delta$. We say that the joint distribution of X is *hierarchical* with underlying simplicial complex Δ (or generating set Δ) if the probability $p(i) = P(X = i)$ of a single cell $i = (i_v, v \in V)$ is of the form

$$\log(p(i)) = \sum_{D \in \Delta} \theta_D(i_D) \tag{1}$$

where $\theta_D(i_D)$ is a function of the marginal cell $i_D = (i_v, v \in D)$ only. The parametrization (1) is not identifiable; that is, for any joint distribution p from the hierarchical model there are different choices for the functions θ_D that satisfy (1). For an example of identifiable and non-identifiable parametrization on the same problem, see the sub-subsection at the end of this subsection. One way to make the parameters unique is to choose a special element within each set I_v, which we denote by 0. The cell with all its components equal to 0 will be denoted by 0 also. The choice of 0 is arbitrary, and a different choice of 0 leads to a simple affine change of parameters. Then, with this choice of a fixed cell 0, by Moebius inversion formula, (1) is equivalent to

$$\theta_E(i_E) = \sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p(i_F, 0_{F^c}) \tag{2}$$

If we need to make precise the dependence of p on θ, then we write $p_\theta(i)$ instead of $p(i)$. The set of all such distributions $\mathcal{E}_\Delta := \{p_\theta\}$ is called the *hierarchical model* of Δ. We now show that $\theta_E(i_E) = 0$ unless all $i_\gamma \neq 0, \gamma \in E$.

Lemma 1.1. *If for $\gamma \in E, E \subseteq V$ we have $i_\gamma = 0$, then $\theta_E(i_E) = 0$*

Proof. By definition and since $(i_{F\cup\gamma}, 0_{(F\cup\gamma)^c}) = (i_F, 0_{F^c})$ if $i_\gamma = 0_\gamma = 0$, we have

$$\begin{aligned}\theta_E(i_E) &= \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, 0_{F^c}) - \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_{F\cup\gamma}, 0_{(F\cup\gamma)^c}) \\ &= \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, 0_{F^c}) - \sum_{F\subseteq E\setminus\gamma} (-1)^{|E\setminus F|} \log p(i_F, 0_{F^c}) = 0\ .\end{aligned}$$

□

From this lemma, it follows immediately that our parametrization is indeed the "baseline" or "corner" constraint parametrization that sets to 0 the values of the $E-$interaction loglinear parameters when at least one index in E is at level 0 – see Agresti (1990), p.150. Therefore, for each $E \subseteq V$, there are only $\prod_{\gamma\in E}(|\mathcal{I}_\gamma| - 1)$ parameters.

Thus, we arrive at the identifiable parametrization

$$\log p_\theta(i) = \theta_0 + \sum_{D\in\Delta\setminus\{\emptyset\}, i_v\neq 0, \forall v\in D} \theta_D(i_D), \tag{3}$$

where $\theta_0 := \theta_\emptyset$. We separate the parameter θ_0 corresponding to the empty set, since it has a special role. It does not depend on the cell index i and acts as a normalizing constant: When all other parameters are chosen freely, θ_0 is determined by the requirement $\sum_{i\in I} p_\theta(i) = 1$. To make it clear that we consider θ_0 as a dependent parameter, we derive it explicitly

$$-\theta_0 = k(\theta) = \log\Big(\sum_{i\in I} \exp\Big(\sum_{D\in\Delta\setminus\{\emptyset\}, i_v\neq 0, \forall v\in D} \theta_D(i_D)\Big)\Big).$$

The parametrization (3) can be further reformulated using the definitions

$$\begin{aligned} S(i) &= \{v \in V\ ;\ i_v \neq 0\} \\ J &= \{j \in I\setminus\{0\},\quad S(j) \in \Delta\}. \end{aligned}$$

For a given $D \in \Delta$ and for given $\theta_D(i_D)$ such that $i_\gamma \neq 0, \forall\gamma \in D$, there is only one $j \in J$ such that $S(j) = D$ and $j_D = j_{S(j)} = i_D$ and conversely. We can therefore write

$$\theta_D(i_D) = \theta_j \ \text{ for the unique } \ j \in J \ \text{ with } \ S(j) = D,\ i_D = j_D.$$

To simplify the notation, we write $j \lhd i$ whenever $S(j) \subseteq S(i)$ and $j_{S(j)} = i_{S(j)}$. Then the parametrization (3) in terms of the *free* parameters $\theta = \{\theta_j,\ j \in J\}$ becomes

$$\log p_\theta(i) = \sum_{j \in J : j \lhd i} \theta_j - k(\theta). \tag{4}$$

It is convenient to introduce the vectors

$$f_i = \sum_{j \in J : j \lhd i} e_j, \quad i \in I$$

where $e_j, j \in J$ are the unit vectors in R^J. Moreover, let A be the $J \times I$ matrix with columns f_i, $i \in I$, and let $\tilde{A}$ be the $(1 + |J|) \times I$ matrix with columns equal to $\binom{1}{f_i}$, $i \in I$. The representation (4) becomes

$$\log p_\theta(i) = \langle \theta, f_i \rangle - k(\theta) \tag{5}$$

$$\log p_\theta = A^t \theta - k(\theta) = \tilde{A}^t \tilde{\theta}, \tag{6}$$

where $\tilde{\theta} = (\theta_0, \theta)$ as a column vector. Both A and $\tilde{A}$ are called *design matrices* of the model.

From the definition of $f_i, i \in I$, it follows immediately that if $n = (n(i), i \in I)$ denotes the I-dimensional column vector of cell counts, then

$$\tilde{A}n = \begin{pmatrix} N \\ t \end{pmatrix} \quad \text{and} \quad An = t, \tag{7}$$

where $N = \sum_{i \in I} n(i)$ is the total cell counts and t is the column vector of $j_{S(j)}$-marginal counts $n(j_{S(j)})$, i.e. $t = (t_j, j \in J)$ where $t_j = n(j_{S(j)}) = \sum_{i | i_{S(j)} = j_{S(j)}} n(i),\ j \in J$.

It follows from (7) that $\frac{t}{N} = \sum_{i \in I} \frac{n(i)}{N} f_i$. Therefore, t belongs to the convex polytope with extreme points $f_i, i \in I$. This polytope is called the *marginal polytope* of the hierarchical model, and we denote it by $\mathbf{P}_\Delta$.

Example 1.2. For the model defined by $V = \{a, b, c\}$, $I_a = \{0, 1\} = I_b = I_c$ and $\Delta = \{a, b, c, ab, bc\}$, we have $I = (000, 100, 010, 110, 001, 101, 011, 111)$ and $J = \{(100), (010), (001), (110), (011)\}$. Then

$$\tilde{A} = \begin{array}{c} \overbrace{}^{f_{000}} \overbrace{}^{f_{001}} \overbrace{}^{f_{010}} \overbrace{}^{f_{011}} \overbrace{}^{f_{100}} \overbrace{}^{f_{101}} \overbrace{}^{f_{110}} \overbrace{}^{f_{111}} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \theta_{000} \\ \theta_{100} \\ \theta_{010} \\ \theta_{001} \\ \theta_{110} \\ \theta_{011} \end{array}$$

An important subclass of hierarchical model is the class of graphical models. Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E. A subset $D \subseteq V$ is a clique of G if any $i, j \in D$, $i \neq j$, define an edge $(i, j) \in E$. The set of cliques of G, denoted by $\Delta(G)$, is a simplicial complex. The *graphical model* of G is defined as the hierarchical model of $\Delta(G)$. Graphical models are important because of their interpretation in terms of conditional independence, see Lauritzen (1996).

1.1.1 An example of identifiable and non-identifiable parametrization

In this section, we will use an example to show that our parameterization is identifiable and compare to another unidentifiable parameterization given by Eriksson(2006).

Example 1.3. For the model defined by $V = \{a, b, c\}$, $\Delta = \{a, b, c, ab, bc\}$ and $I_a = \{0, 1\} = I_b = I_c$, we have $I = \{000, 100, 010, 110, 001, 101, 011, 111\}$ and $J = \{(100), (010), (001), (110), (011)\}$. Then

$$\tilde{A} = \begin{array}{c} \begin{matrix} \overbrace{}^{f_{000}} & \overbrace{}^{f_{001}} & \overbrace{}^{f_{010}} & \overbrace{}^{f_{011}} & \overbrace{}^{f_{100}} & \overbrace{}^{f_{101}} & \overbrace{}^{f_{110}} & \overbrace{}^{f_{111}} \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \begin{matrix} \theta_{000} \\ \theta_{001} \\ \theta_{010} \\ \theta_{100} \\ \theta_{011} \\ \theta_{110} \end{matrix}$$

Since θ_0 is not a free parameter in our setting, and matrix A is a full rank matrix, our parameterization is identifiable. To verify this, we can extract the columns indexed by the set J, rows indexed by the set θ, and get the following sub-matrix:

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

The inverse of X is

$$X^{-1} = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Therefore we have

$$\theta = (X^{-1})^t \log \frac{p_\theta}{p_\theta(0)}$$

$$\begin{pmatrix} \theta_{001} \\ \theta_{010} \\ \theta_{100} \\ \theta_{011} \\ \theta_{110} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \log \frac{p(001)}{p(000)} \\ \log \frac{p(010)}{p(000)} \\ \log \frac{p(011)}{p(000)} \\ \log \frac{p(100)}{p(000)} \\ \log \frac{p(110)}{p(000)} \end{pmatrix}$$

It's easy to verify that the above equation system gives us the formulas of θ as we defined before.

Now let's check another parameterization given by Eriksson(2006). Their design matrix A_Δ is a 0-1 matrix whose rows are indexed by the facets of simplicial complex Δ taking different values and columns are indexed by I. In this example, their parameters are

$$\{\theta_{ab}(00), \theta_{ab}(10), \theta_{ab}(01), \theta_{ab}(11), \theta_{bc}(00), \theta_{bc}(10), \theta_{bc}(01), \theta_{bc}(11)\}.$$

The design matrix is given as follows:

$$A_\Delta = \begin{array}{c} \begin{matrix} \overbrace{}^{f_{000}} & \overbrace{}^{f_{001}} & \overbrace{}^{f_{010}} & \overbrace{}^{f_{011}} & \overbrace{}^{f_{100}} & \overbrace{}^{f_{101}} & \overbrace{}^{f_{110}} & \overbrace{}^{f_{111}} \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \theta_{ab}(00) \\ \theta_{ab}(10) \\ \theta_{ab}(01) \\ \theta_{ab}(11) \\ \theta_{bc}(00) \\ \theta_{bc}(10) \\ \theta_{bc}(01) \\ \theta_{bc}(11) \end{matrix}$$

The rank of A_Δ is 6, i.e. it is not a full rank matrix. We can't solve $\log p = A_\Delta^t \theta$ to get the formulas of θ as A_Δ is noninvertible. This parameterization is not identifiable. Denote r_i as the i-th row of A_Δ, we can see that $r_1 + r_2 = r_5 + r_7$, and $r_3 + r_4 = r_6 + r_8$.

1.2 Discrete exponential families

Hierarchical models are examples of discrete exponential families. Let I and J be finite sets, and let $A \in \mathbf{R}^{J \times I}$ be a real matrix. Denote the columns of A by f_i, $i \in I$. The discrete exponential family corresponding to A, denoted by $\mathcal{E}_A$, consists of all probability distributions on I that are of the form

$$p_\theta(i) = \exp(\langle \theta, f_i \rangle - k(\theta)), \qquad \theta \in \mathbf{R}^J,$$

where $k(\theta) = \log \sum_{i \in I} \exp(\langle \theta, f_i \rangle)$. Define $\tilde{J} = J \cup \{0\}$, $\theta_0 = -k(\theta)$ and $\tilde{\theta} = (\theta_0, \theta)$, and let $\tilde{A}$ be the matrix A with one additional row of ones; that is, $\tilde{A} \in \mathbf{R}^{\tilde{J} \times I}$ is the matrix with columns $\binom{1}{f_i}$, $i \in I$. We make the additional assumption that the row of 1's belongs to the row span of the matrix A so that A and $\tilde{A}$ have the same rank. Then $\mathcal{E}_A$ consists of the probability distributions p_θ that satisfy $\log p_\theta = \tilde{A}^t \tilde{\theta}$ for some $\tilde{\theta} \in \mathbf{R}^{\tilde{J}}$. The convex hull of the columns f_i, $i \in I$, is called the *convex support polytope*, denoted by $\mathbf{P}_A$. It generalizes the marginal polytope.

The parametrization $\theta \to p_\theta$ is identifiable if and only if A has full rank. If A does not have full rank, then one can drop certain rows of A to obtain a submatrix A' with full rank. This is equivalent to setting certain parameters to zero until the remaining parameters are identifiable.

For a discrete exponential family, $\prod_{i \in I} p_\theta(i)^{n(i)}$ can be written under the form of a natural exponential family. Indeed,

$$\begin{aligned}\prod_{i \in I} p_\theta(i)^{n(i)} = \exp(\sum_{i \in I} n(i) \log p_\theta(i)) = \exp(\langle n, \log(p_\theta) \rangle) = \exp(\langle n, \tilde{A}^t \tilde{\theta} \rangle) \\ = \exp(\langle \tilde{A} n, \tilde{\theta} \rangle) = \exp(\sum_{j \in J} \theta_j t_j - N k(\theta)).\end{aligned}$$

The log-likelihood function in θ for the loglinear parameter of the multinomial distribution with the given hierarchical model is therefore

$$l(\theta) = \sum_{j \in J} \theta_j t_j - N k(\theta). \tag{8}$$

It is well-known that $l(\theta)$ is concave. If the parameters are identifiable, then it is strictly concave.

We can also express the log-likelihood as a function of $\mu = (\mu_i = \log \frac{p(i)}{p(0)}, i \in I)$:

$$l(\mu) = \sum_{i \in I} n(i) \log p(i) = \sum_{i \in I \setminus \{0\}} n(i) \log \frac{p(i)}{p(0)} + N \log p(0)$$
$$= \sum_{i \in I} n(i)\mu_i - N \log(\sum_{i \in I} \exp \mu_i). \quad (9)$$

As stated before, only a subset μ_L of the parameters μ are independent, and the remaining μ_i, $i \notin L$, can be expressed as linear functions of μ_L.

1.3 Polytopes

We next recall some general facts about polytopes and their faces. We refer to Ziegler (1998), Lectures on Polytopes, for details and more information.

Definition 1.4. A set $\mathbf{P} \subset \mathbf{R}^h$ is a *(convex) polytope* if $\mathbf{P}$ is the convex hull of a finite subset of $\mathbf{R}^h$. Equivalently, a polytope can be defined as a bounded subset of $\mathbf{R}^h$ defined by linear inequalities.

Definition 1.5. For any vector $g \in \mathbf{R}^h$ and any constant $c \in \mathbf{R}$, define three sets $H_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle = c\}$, $H^+_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle \geq c\}$ and $H^-_{g,c} = \{x \in \mathbf{R}^h : \langle g, x \rangle \leq c\}$. If $g \neq 0$, then $H_{g,c}$ is an *(affine) hyperplane*, and $H^+_{g,c}$ and $H^-_{g,c}$ are the *positive* and *negative halfspace* defined by g and c.

Let $\mathbf{P} \subseteq \mathbf{R}^h$ be a polytope, let $g \in \mathbf{R}^h$ and $c \in \mathbf{R}$, and suppose that $\mathbf{P} \subset H^+_{g,c}$ or $\mathbf{P} \subset H^-_{g,c}$. Then $\mathbf{F} := H_{g,c} \cap \mathbf{P}$ is called a *face* of $\mathbf{P}$. If $g \neq 0$, then $H_{g,c}$ is called a *supporting hyperplane* of $\mathbf{P}$. If $\mathbf{F} \neq \mathbf{P}$ and $\mathbf{F} \neq \emptyset$, then $\mathbf{F}$ is a *proper face* of $\mathbf{P}$.

The dimension of a face $\mathbf{F}$ is the dimension of the smallest affine subspace of $\mathbf{R}^h$ that contains it. Its co-dimension is $\dim(\mathbf{P}) - \dim(\mathbf{F})$. A *facet* of a polytope $\mathbf{P}$ is a proper face that is maximal with respect to inclusion and is thus of co-dimension 1. A minimal proper face of a polytope is a singleton $\{p\} \subseteq \mathbf{P}$; in this case, p is a *vertex*.

Intersections of faces are again faces: If $g_1, g_2 \in \mathbf{R}^h$ and $c_1, c_2 \in \mathbf{R}$ define faces $\mathbf{F}_1, \mathbf{F}_2$ of $\mathbf{P}$ and if $\mathbf{P} \subset H^+_{g_1,c_1} \cap H^+_{g_2,c_2}$, then $\mathbf{P} \subset H^+_{g_1+g_2,c_1+c_2}$, and $\mathbf{F}_1 \cap \mathbf{F}_2 = \mathbf{P} \cap H_{g_1+g_2,c_1+c_2}$ (if x is such that $\langle g_1, x \rangle = c_1$ and $\langle g_2, x \rangle = c_2$, then clearly that x must be such that $\langle g_1 + g_2, x \rangle = c_1 + c_2$). Any face is an intersection of facets.

By definition, every face $\mathbf{F}$ of a polytope $\mathbf{P} \subset \mathbf{R}^h$ is characterized by a linear inequality $\langle g, x \rangle \geq c$ that is valid on $\mathbf{P}$ and that holds as an equality on $\mathbf{F}$. This linear inequality is unique only if $\mathbf{F}$ is a facet. Sometimes it is convenient to give all linear equations that hold on a face $\mathbf{F}$. These linear equations determine the smallest affine subspace of $\mathbf{R}^h$ containing $\mathbf{F}$.

When a polytope is defined as the convex hull of a finite number of points f_i, $i \in I$, then it is of interest to know, which subsets of $\{f_i\}_{i \in I}$ lie on a common face. Indeed, it is the purpose of our work to compute the smallest face of the marginal polytope containing the data vector t, and we will determine this face by identifying which vectors f_i belong to it.

Definition 1.6. For a finite set I let $\{f_i\}_{i \in I} \subset \mathbf{R}^h$, and let $\mathbf{P}$ be the convex hull of $\{f_i\}_{i \in I}$. A subset $F \subseteq I$ is called *facial* (with respect to $\mathbf{P}$), if there exists a face $\mathbf{F}$ of $\mathbf{P}$ with $F = \{i : f_i \in \mathbf{F}\}$. For any subset $S \subseteq I$, denote by $F_{\mathbf{P}}(S)$ the smallest facial set that contains S.

Since the intersection of facial sets is again facial, $F_{\mathbf{P}}(S)$ is well-defined.

Lemma 1.7. *Let $\{f_i\}_{i \in I} \subset \mathbf{R}^h$, let $\phi : \mathbf{R}^h \to \mathbf{R}^{h'}, x \mapsto Bx + d$ be an affine map, and let $f'_i = \phi(f_i)$. If $\mathbf{P}$ is the convex hull of the f_i, then $\mathbf{P}' := \phi(\mathbf{P})$ is the convex hull of the f'_i. The faces and facial sets of $\mathbf{P}$ and $\mathbf{P}'$ are related as follows:*

1. *Any inequality $\langle g', x' \rangle \geq c'$ that is valid on $\mathbf{P}'$ corresponds to an inequality $\langle g, x \rangle \geq c$ that is valid on $\mathbf{P}$, where $g = B^t g'$ and $c = c' - \langle g', d \rangle$. Thus, if $\mathbf{F}'$ is a face of $\mathbf{P}'$, then $\phi^{-1}(\mathbf{F}')$ is a face of $\mathbf{P}$.*

2. *A subset of I that is facial with respect to $\mathbf{P}'$ is also facial with respect to $\mathbf{P}$. Thus, $F_{\mathbf{P}}(S) \subseteq F_{\mathbf{P}'}(S)$ for any $S \subseteq I$.*

Proof. The first statement follows from

$$c' \leq \langle g', \phi(f_i) \rangle = \langle g', Bf_i + d \rangle = \langle B^t g', f_i \rangle + \langle g', d \rangle,$$

which holds for any $i \in I$. We note that if ϕ is a simple projection , then g is obtained from g' by adding 0 components to the additional dimensions.

For the second statement, since the lift of a face in $\mathbf{P}'$ is also a face in $\mathbf{P}$, the f'_i in the face of $\mathbf{P}'$ are lifted to f_i in $\mathbf{P}$ and define a face (because any $x' \in \mathbf{P}'$ is convex combination of the f'_i in $\mathbf{P}'$ and thus the lift of x' are linear combinations of the lift of f'_i in the face of $\mathbf{P}'$. Thus a subset of I that is

facial with respect to $\mathbf{P}'$ is also facial with respect to $\mathbf{P}$. For the last sentence of the second statement, we have from what we just proved that $F_{\mathbf{P}'}(S)$ is facial with respect to $\mathbf{P}$ and also it contains S. Thus $F_{\mathbf{P}}(S)$ which is the smallest facial set with respect to $\mathbf{P}$ and that contains S must be included in $F_{\mathbf{P}'}(S)$. □

We note that in Lemma 1.7, the dimension of $\phi(\mathbf{P})$ is at most equal to h. We will only apply Lemma 1.7 to coordinate projections ϕ with $h' < h$.

Remark 1.8. Sometimes it is convenient to embed the polytope in a vector space that has one additional dimension using a map $\mathbf{R}^h \to \mathbf{R}^{h+1}, x \mapsto \tilde{x} := (1, x)$. This has the advantage that all defining inequalities can be brought into a homogeneous form with vanishing constant c: Note that $\langle g, f_i \rangle - c = \langle \tilde{g}_c, \tilde{f}_i \rangle$, where $\tilde{g}_c := (c, g)$.

When a defining inequality of a face $\mathbf{F}$ is given, its facial set F can be obtained by checking whether $f_i \in \mathbf{F}$ for each $i \in I$. In the other direction, when a facial set F is given, it is much more difficult to compute a defining inequality of the corresponding face $\mathbf{F}$. However, it is straightforward to compute the linear equations defining $\mathbf{F}$: the set of such equations $0 = \langle g, x \rangle - c = \langle \tilde{g}, \tilde{x} \rangle$ corresponds to the set of vectors $\tilde{g} \in \ker \tilde{A}_F^t$, where $\tilde{A}_F$ is the matrix obtained from A by adding a row of ones and dropping the columns not in F.

1.4 The closure of an exponential family and existence of the mle

We fix a discrete exponential family $\mathcal{E}_A$. While our main interest lies in hierarchical models, the results that we need are more naturally formulated in the language of discrete exponential families. We assume that a vector of observed counts $n = (n(i) : i \in I)$ is given.

Definition 1.9. A parameter value θ^* is a *maximum likelihood estimate* (mle) if it is a global maximum of $l(\theta)$.

The function $l(\theta)$ is always bounded (clearly, it is never positive). The fundtion $l(\theta)$ is strictly concave, and so the maximum is unique, if it exists. However, a maximum need not exist, since the domain of the parameters θ is unbounded. To understand this, it is convenient to interpret the likelihood

as a function of probabilities. Let $\tilde{l}$ be the function that assigns to any probability distribution p on I the value

$$\tilde{l}(p) = \log(\prod\nolimits_{i\in I} p(i)^{n(i)})$$

Then $l(\theta) = \tilde{l}(p_\theta)$, and θ^* is an mle if and only if p_{θ^*} maximizes $\tilde{l}$ subject to the constraint that p belongs to the hierarchical model (and thus is of the form p_θ for some θ). While the set of all probability distributions on I is compact, the hierarchical model itself is not closed (indeed, from the definition of $\mathcal{E}_A$, the various $p(i)$ cannot be equal to 0 because $p(i) = \frac{\exp\langle\theta,f_i\rangle}{\sum_{j\in I}\exp\langle\theta,f_j\rangle}$. If one of the components θ_α of θ goes to $+\infty$ where $f_{i,\alpha} \neq 0$, then $p(i) = \frac{\exp(\theta_\alpha f_{i,\alpha}+\sum_{\beta\neq\alpha}\theta_\beta f_{i,\beta})}{\exp(\theta_\alpha f_{i,\alpha}+\sum_{\beta\neq\alpha}\theta_\beta f_{i,\beta})+\sum_{j\in I, j\neq i}\exp\langle\theta,f_j\rangle}$ will go to 1 and the others will have to go to 0 and thus this $p = (p(i), i \in I)$ will not be in the hierarchical model)) and therefore not compact, and so there is no guarantee that $\tilde{l}$ attains its maximum on the hierarchical model. However, things become better when we pass from the hierarchical model to its topological closure, where the topology comes from interpreting a probability distibution as a vector $p = (p(i))_{i\in I} \in \mathbf{R}^I$ of real numbers (this choice of the topology is canonical since we are dealing with a finite set I. The closure is sometimes also called *completion* (see Barndorff:exponential families, (1978). Since the closure of the hierarchical model is again compact, the continuous function $\tilde{l}$ always attains its maximum.

Theorem 1.10. *The closure of a discrete exponential family can be written as a union*

$$\overline{\mathcal{E}_A} = \bigcup_F \mathcal{E}_{F,A},$$

where F runs over all facial sets of the convex support polytope $\mathbf{P}_A$ and where $\mathcal{E}_{F,A}$ consists of all probability distributions of the form $p_{F,\theta}$, with

$$p_{F,\theta} = \begin{cases} \exp(\langle\theta, f_i\rangle - k_F(\theta)), & \text{if } i \in F, \\ 0, & \text{otherwise,} \end{cases}$$

where $k_F(\theta) = \log\sum_{i\in F}\exp(\langle\theta, f_i\rangle$.

Proof. See Barndorff-Nielsen (1976). For self-containedness we provide a proof in our notation in Appendix A.1. □

Theorem 1.10 shows that $\overline{\mathcal{E}_A}$ is a finite union of sets $\mathcal{E}_{F,A}$ that are exponential families themselves with a very similar parametrization, using the same number of parameters and the same design matrix A (or, rather, the submatrix A_F consisting of those columns of A indexed by F). However, for any proper facial set F, the parametrization $\theta \mapsto p_{F,\theta}$ is not injective, i.e. the parameters θ are not identifiable on $\mathcal{E}_{F,\Delta}$. The reason is that the matrix $\tilde{A}_F$ does not have full rank, even if $\tilde{A}$ has full rank, since all columns of $\tilde{A}_F$ lie on a supporting hyperplane defining F.

A second thing to note is that although the parameters θ on $\mathcal{E}_A$ and the parameters θ on $\mathcal{E}_{F,A}$ play similar roles, they are very different in the following sense: If $\theta^{(s)}$ is a sequence of parameters with $p_{\theta^{(s)}} \to p_{F,\theta}$ for some θ, then, in general, $\lim_{s\to\infty} \theta_j^{(s)} \neq \theta_j$ for all $j \in J$.

Theorem 1.11. *For any vector of observed counts n, there is a unique maximum p^* of $\tilde{l}$ in $\overline{\mathcal{E}_A}$. For t as defined in (5), this maximum p^* satisfies:*

- $Ap^* = \frac{t}{N}$.
- $\operatorname{supp}(p^*) = F_t$.

Proof. See Barndorff-Nielsen (1978). For self-containedness we provide a proof in our notation in Appendix A.2. □

Definition 1.12. The maximum in Theorem 1.11 is called the *extended* maximum likelihood estimate (EMLE).

Clearly, if the mle θ^* exists, then $p^* = p_{\theta^*}$.

2 A Bayesian perspective

2.1 The DY conjugate prior

We saw that for a discrete exponential family, $\prod_{i\in I} p_\theta(i)^{n(i)}$ can be written under the form of a natural exponential family. Indeed,

$$\begin{aligned}\prod_{i\in I} p_\theta(i)^{n(i)} &= \exp(\sum_{i\in I} n(i)\log p_\theta(i)) = \exp(\langle n, \log(p_\theta)\rangle) = \exp(\langle n, \tilde{A}^t\tilde{\theta}\rangle) \\ &= \exp(\langle \tilde{A}n, \tilde{\theta}\rangle) = \exp(\sum_{j\in J}\theta_j t_j - Nk(\theta)). \qquad (10)\end{aligned}$$

For a contingency table $X = (X_v, v \in V)$ is the random variable. We can then write

$$f(X) = \exp(\sum_{j \in J} \theta_j t_j(X) - Nk(\theta)) = \exp(\langle \theta, t \rangle - Nk(\theta)).$$

From the form (10) of the multinomial distribution and Theorem 1 inDiaconis-Ylvisaker (1979), the DY conjugate prior distribution for θ has density with respect to the Lebesgue measure equal to

$$\pi(\theta | m_J, \alpha, J) = \frac{1}{I_J(m_J, \alpha)} \times \frac{e^{\alpha \langle \theta, m_J \rangle}}{L(\theta)^\alpha} = \frac{1}{I_J(m_J, \alpha)} \times e^{\alpha \langle \theta, m_J \rangle - \alpha k(\theta)}$$

where $I_J(m, \alpha)$ is the normalizing constant. It is proper if and only if the hyperparameter (α, m_J) is such $\alpha > 0$ and $m_J \in C$. The posterior probability of θ given the data $n = (n(i))_{i \in I}$ and $t_J = (t_j, j \in J)$ is

$$\pi(\theta | \frac{\alpha m_J + t_J}{\alpha + N}, \alpha + N, J).$$

In classical Bayesian model selection, the most probable models are selected by means of Bayes factors. More precisely, models are compared two by two by means of the Bayes factor $B_{1,2}$ between model J_1 and model J_2. If the prior on the set of all hierarchical models is uniform, we have

$$B_{1,2} = \frac{I_2(m_2, \alpha)}{I_1(m_1, \alpha)} \times \frac{I_1(\frac{\alpha m_1 + t_1}{\alpha + N}, \alpha + N)}{I_2(\frac{\alpha m_2 + t_2}{\alpha + N}, \alpha + N)} \tag{11}$$

where, for the sake of simplicity, m, t, I are indexed by $k = 1, 2$ rather than by J_1, J_2 and where m_1 and m_2 have been chosen in C_1 and C_2 respectively. The aim of the present paper is to find the limit of $B_{1,2}$ when $\alpha \to 0$. If we assume that $n(i) > 0$ for all $i \in I$, then t_k/N is in the interior of C_k and under these circumstances the second factor in the right-hand side of (11) has the finite limit $I_1(\frac{t_1}{N}, N)/I_2(\frac{t_2}{N}, N)$. For the first factor in (11), we will show that $I(m, \alpha) \sim_{\alpha \to 0} \mathbb{J}_C(m) \alpha^{-|J|}$ where $\mathbb{J}_C(m)$ will be defined in the next section. Thus when $\alpha \to 0$ the Bayes factor is equivalent to

$$\alpha^{|J_1| - |J_2|} \frac{\mathbb{J}_{C_2}(m_2)}{\mathbb{J}_{C_1}(m_1)} \times \frac{I_1(\frac{t_1}{N}, N)}{I_2(\frac{t_2}{N}, N)}.$$

If we do not assume that $n(i) > 0$ for all $i \in I$, then t_k/N might be on the boundary of C_k for at least one $k = 1, 2$ and we will have to further study the behaviour of $I(m, \alpha)$ and $\mathbb{J}_C(m)$. This is done in the following section.

2.2 The characteristic function of a convex set

If $C \subset E$ is an open non empty convex set not containing an (affine) line,its support function $h_C : E^* \to (-\infty, \infty]$ is

$$h_C(\theta) = \sup\{\langle \theta, x\rangle \ ; \ x \in C\}$$

and its characteristic function is the function $m \mapsto \mathbb{J}_C(m)$ defined on C by

$$\mathbb{J}_C(m) = \int_{E^*} e^{\langle \theta, m\rangle - h_C(\theta)} d\theta. \tag{12}$$

We note that if C contained a line, we would have $h_C(\theta) = \infty$ almost everywhere and $\mathbb{J}_C \equiv 0$. Faraut and Koranyi (1994), p. 10 define $\mathbb{J}_C$ when C is an open convex salient cone. In that case, the polar set of C is the convex cone

$$C^o = \{\theta \in E^* \ ; \ \langle \theta, x\rangle \leq 0 \ \forall x \in C\} \tag{13}$$

and $h_C(\theta) = 0$ if $\theta \in C^o$ and $h_C(\theta) = \infty$ if $\theta \notin C^o$. When C is a bounded set, $h_C(\theta)$ is finite for all $\theta \in E^*$. We also have the following important property of $\mathbb{J}_C(.)$.

Lemma 2.1. *Let C be an open convex set not containing a line and let $m \in C$. Then $\mathbb{J}_C(m)$ is finite.*

Example 2.2. Let $C = (0, 1) \subseteq \mathbb{R}$. In this case, $h_C(\theta) = \max(0, \theta)$ and for $0 < m < 1$ we have

$$\mathbb{J}_C(m) = \int_{-\infty}^{0} e^{\theta m} d\theta + \int_0^{\infty} e^{\theta m - \theta} d\theta = \frac{1}{m} + \frac{1}{1-m} = \frac{1}{m(1-m)}. \tag{14}$$

Theorem 2.3. *Let $C \subset E$ be the non empty interior of a bounded polytope $\overline{C}$. Let $m \in C$. Then we have*

$$\mathbb{J}_C(m) = \frac{N(m)}{D(m)}$$

where $D(m) = \prod_{k=1}^{K} g_k(m)$ is the product of affine forms $g_k(m)$ in m such that $g_k(m) = 0, \ k = 1, \ldots, K$ define the facets of $\overline{C}$ and where $N(m)$ is a polynomial of degree $< K$.

Theorem 2.4. *Let $C \subset E$ be an open polytope with $\dim E = n$. Let $y \in \partial C$, let F be the face of $\overline{C}$ containing y in its relative interior and let k be the dimension of F. Then when $\lambda \to 0$*

$$\lim_{\lambda \to 0} \lambda^{n-k} \mathbb{J}_C(\lambda m + (1-\lambda) y) = D,$$

where D is a positive constant.

2.3 The behaviour of $I(m,\alpha)$ as $\alpha \to 0$

We have the following theorem.

Theorem 2.5. *Let μ be a positive measure on the n-dimensional linear space E with closed convex support bounded and with nonempty interior C. Denote by $L(\theta) = \int_E e^{\langle \theta, x\rangle}\mu(dx)$ its Laplace transform. For $m \in C$ and for $\alpha > 0$ consider the Diaconis Ylvisaker integral*

$$I(m,\alpha) = \int_{E^*} \frac{e^{\alpha\langle \theta, m\rangle}}{L(\theta)^\alpha} d\theta.$$

Then

$$\lim_{\alpha \to 0} \alpha^n I(m,\alpha) = \mathbb{J}_C(m). \tag{15}$$

Let us note immediately that a remarkable feature of this result is that the limit $\mathbb{J}_C(m)$ of $\alpha^n I(m,\alpha)$ depends on μ only through its convex support. For instance if $E = \mathbb{R}$, the uniform measure on $(0,1)$ and the sum $\mu = \delta_0 + \delta_1$ of two Dirac measures share the same $C = (0,1)$ and the same $\mathbb{J}_C(m) = (m(1-m))^{-1}$.

Lemma 2.6. *Let μ be a bounded measure on some measurable space Ω and let f be a positive, bounded and measurable function on Ω. Then we have*

1. $||f||_p \to_{p\to\infty} ||f||_\infty$
2. *The function $p \mapsto ||f||_p$ is either decreasing on $(0,\infty)$ or there exists $p_0 \geq 0$ such that it is decreasing on $(0,p_0]$ and increasing on $[p_0,+\infty)$.*

Proof. (OF THEOREM 2.5.) In the integral $\alpha^n I(m,\alpha)$ we make the change of variable $y = \alpha\theta$ and we obtain

$$\alpha^n I(m,\alpha) = \int_{E^*} \frac{e^{\langle y, m\rangle}}{L(y/\alpha)^\alpha} dy.$$

We now apply the last lemma to $\Omega = \overline{C}$, to the bounded measure μ, to the function $f(x) = e^{\langle y,x\rangle}$ for some fixed $y \in E^*$ and to $p = 1/\alpha$. Denote by S the support of μ. One easily sees that the support function of C satisfies

$$h_C(\theta) = \sup\{\langle \theta, x\rangle \ ; \ x \in C\} = \max\{\langle \theta, x\rangle \ ; \ x \in S\}$$

since C is the interior of the convex hull of S. As a consequence the essential sup of f is $e^{h_C(y)}$ and we get $\lim_{\alpha\to 0} L(y/\alpha)^\alpha = e^{h_C(y)}$. Furthermore, by Lemma 2.6, the function $p \mapsto \|f\|_p$ is monotonic for p big enough. If $p \mapsto ||f||_p$ is increasing, $\frac{1}{||f||_p}$ is decreasing and then by the monotone convergence theorem

$$\lim_{\alpha\to 0} \int_{E^*} \frac{e^{\langle y,m\rangle}}{L(y/\alpha)^\alpha} dy \quad = \quad \int_{E^*} \frac{e^{\langle y,m\rangle}}{\lim_{\alpha\to 0} L(y/\alpha)^\alpha} dy = \int_{E^*} e^{\langle y,m\rangle - h_C(y)} dy = \mathbb{J}_C(m)$$

If $p \mapsto ||f||_p$ is decreasing, $p \mapsto 1/||f||_p$ is increasing. In order to show that we can invert the order of limit and integration and to apply the monotone convergence theorem as we did in the previous case, we need to insure that $\int_{E^*} e^{\langle y,m\rangle - h_C(y)} dy$ is finite: Lemma 2.1 shows that it is true. □

2.4 The case where the data belongs to a face of $\overline{C}_i$, $i = 1, 2$

When $\alpha \to 0$, $\frac{\alpha m_i + t_i}{\alpha + N}$ converges to the boundary point $\frac{t_i}{N}$ of C_i along the segment

$$s(\alpha) = \frac{\alpha m_i + t_i}{\alpha + N} = \frac{\alpha}{\alpha + N} m_i + (1 - \frac{\alpha}{\alpha + N}) \frac{t_i}{N}. \tag{16}$$

We need to study the limiting behaviour of $B_{1,2}$ when $\alpha \to 0$. To do so, we will use Theorem 2.4 to obtain the following result.

Theorem 2.7. *Suppose that $\frac{t}{N} \in \overline{C} \setminus C$ belongs to the relative interior of a face F of dimension k. Then*

$$\lim_{\alpha\to 0} \alpha^{(|J|-k)} I(\frac{\alpha m + t}{\alpha + N}, \alpha + N) \tag{17}$$

exists and is positive.

From Theorems 2.5 and 2.7, we immediately derive the following which is the object of this section.

Corollary 2.8. *Consider two hierarchical models J_i, $i = 1, 2$ of dimension $|J_i|$. Assume that the data $\frac{t_i}{N}$ belongs to the relative interior of a face F_i of C_i of dimension k_i. Then the asymptotic behaviour of the Bayes factor $B_{1,2}$ when $\alpha \to 0$ is given by*

$$B_{1,2} \quad \sim \quad D\alpha^{k_1 - k_2}$$

where D is a finite positive constant. The Bayes factor favours the model which contains the data in the relative interior of the face of C_i of smallest dimension.

The proof is immediate. According to Theorems 2.5 and 2.7, we have

$$B_{1,2} = \frac{I(m_2,\alpha)}{I(m_1,\alpha)} \frac{I(\frac{\alpha m_1+t_1}{\alpha+N}, \alpha+N)}{I(\frac{\alpha m_2+t_2}{\alpha+N}, \alpha+N)} \sim \alpha^{|J_1|-|J_2|} \alpha^{(k_1-|J_1|)-(k_2-|J_2|)} = \alpha^{k_1-k_2} .$$

Remark 2.9. We note that, if $\frac{t_i}{N} \in C_i$, $i = 1, 2$, since C_i is the face of $\overline{C_i}$ of dimension J_i, then $k_i = |J_i|$ and Corollary 2.8 yields Corollary that the Bayes factor chooses the sparsest model, following conventional wisdom. For the same reason, Corollary 2.8 also deals with the cases where $\frac{t_i}{N} \in C_i$ for only one of $i = 1$ or $i = 2$.

2.5 Conclusion

We see from the rsults above that the asymptotic behavious of the Bayes factor depnds not on the two models considered but on the face of the models considered. In other words, as far as the Bayes factor is concerned, we have to consider the closed exponential family given by the model and in that family choose the exponential family for which the data belongs to the interior of the marginal cone.

3 Approximating the faces of the marginal polytope

3.1 Condition for the existence of the mle: an algorithm

Recall that the multinomial density can be written under exponential family form as

$$\exp(\langle An, \theta\rangle - Nk(\theta))$$

where the columns of A are the vectors $f_i = \sum_{j \lhd i} e_j$, $i \in I$.

Definition 3.1. We say that the mle does not exist or is undefined if the supremum of the loglikelihood function is not attained by a finite vector θ.

Haberman (1974) was the first to give necessary and sufficient condition for the existence of the mle.

Theorem 3.2. *A necessary and sufficient condition for the existence of the mle is that there exists a vector* $z \in R^I$, $z \in Ker(A)$ *such that* $z + n > 0$.

This can be given a geometric interpretation in terms of the marginal polytope(see Eriksson, Fienberg, Rinaldo and Sullivant (2006)). Very simply $t = An$ and since $z \in Ker(A), A(z + n) = t$ also. Thus, since all the components of $z+n$ are strictly positive, $z+n$ belongs to the relative interior of the marginal polytope and the mle exists.

Corollary 3.3. *The mle for the mean vector* p *exists if and only if the margins* $t = An$ *belongs to the relative interior of the marginal polytope* $\mathbf{P}_A$.

Indeed a vector t belongs to the interior of the marginal polytope if and only if there exists $x > 0, x \in R^I$ such that $t = Ax$. Then Theorem 3.2 states that the mle exists if and only if t belong to the relative interior of $\mathbf{P}_A$: Indeed, $t = An = A(n + z)$ with $n + z > 0$. Thus the mle does not exist if and only if t lies on one of the facets of $\mathbf{P}_A$. Hence, to show t belongs to a facet, we want to show that there exists $g \in (R^{|J|})^*, g \in \mathbf{P}_A^*$ which attains its maximum at t but does not attain it in another other point of $\mathbf{P}_A$. This can be decided by determining if the polyhedral cone

$$\{g \mid g^t A \leq \mathbf{1}^t \cdot g^t t\} \tag{18}$$

contains only those vectors orthogonal to the span of $f_i, i \in I$.

A Linear Programming algorithm to compute facial sets

Denote A as the design matrix, A_+ as the sub-matrix with columns indexed by the positive cells and A_0 as the sub-matrix indexed by the empty cells.

Lemma 3.4. *Solution* g^* *of the non-linear problem*

$$\begin{array}{c} \max_{\tilde{g}} \quad z = \quad \|\tilde{A}^t \tilde{g}\|_0 \\ s.t.\ \tilde{A}_+^t \tilde{g} = 0 \\ \tilde{A}_0^t \tilde{g} \geqslant 0 \end{array} \tag{19}$$

is a perpendicular vector to the smallest face containing t. *The corresponding facial set is* $F_t = I \setminus supp(A^t g^*)$.

Proof. Let us first note that $f_i^t g = c$ for some constant c can be written as $\tilde{f}_i^t \tilde{g} = 0$ where $\tilde{f}_i = (1, f_i^t)^t$ and $\tilde{g} = (-c, g^t)^t$ and thus we can write (18) in a homogeneous way

$$\{\tilde{g} \mid \tilde{g}^t \tilde{A} \geq \tilde{\mathbf{1}}^t \cdot 0\} \,. \tag{20}$$

Since $\tilde{A}^t \tilde{g}$ is the $|I|$-dimensional vector with component $\tilde{f}_i^t \tilde{g}$, we see that $\|\tilde{A}^t \tilde{g}\|_0$ is the number of $i \in I$ such that $\tilde{f}_i^t \tilde{g} > 0$, i.e. the number of f_i that do not belong to the facet containing t. So, by solving the optimization problem (19), we find a supporting hyperplane which contains t and identifies all the f_i's that are not in the smallest face of $\mathbf{P}_A$ containing t. □

The optimization problem (19) is highly non-linear and non-convex: it can be solved by repeatedly solving the associated ℓ_1-norm optimization problem:

$$\begin{aligned} \max \quad & z = \|\tilde{A}_0^t \tilde{g}\|_1 \\ s.t. \quad & \tilde{A}_+^t \tilde{g} = 0 \\ & \tilde{A}_0^t \tilde{g} \geq 0 \\ & \tilde{A}_0^t \tilde{g} \leq 1 \end{aligned} \tag{21}$$

Problem (21) is a linear programming problem: we can solve it repeatedly until we get the smallest facial set F_t. The process is as follows:

Algorithm 1 Face computation by linear programming method

Require: Design matrix A and positive cell index I_+
  INITIALIZE $A_+ = A(:, I_+)$, $A_0 = A \setminus A_+$
  Solve problem 21, get the solution g^* and the corresponding maximum z^*
  while $A_0 \neq \emptyset$ and $z^* \neq 0$ **do**
    Let matrix B be the submatrix of A_0, by taking columns of A_0 which satisfy $\langle f_i, g^* \rangle > 0$, update $A_0 = A_0 \setminus B$,
    Solve problem 21, get the solution g^* and the corresponding maximum z^*
  end while
  if $A_0 = \emptyset$ **then**
    $F_t = I_+$
  end if
  if $Z^* = 0$ **then**
    $F_t = I_+ \cup \{i | i$ is the index of $A_0\}$
  end if

The algorithm will work for problems with up to 16 variables but not more. The question is what to do when we have more than 16 variables and we want to have an idea of whether the mle exists and which $\hat{p(i)}$ are equal to 0.

3.2 Approximating the smallest face containing t

3.2.1 The basic facts

Here are the basic properties that help us build the inner and outer approximations to $\mathbf{F}_t$.

Fact 1: $\mathbf{F}_t$ contains I_+. In fact $\mathbf{F}_t = F_\Delta(I_+)$.

Lemma 3.5. *The sufficient statistic t belongs to the face F_g of C governed by g if and only if $f_i \in F_g$ for all $i \in I_+$.*

The proof is obvious if we write that $t \in F_g \Leftrightarrow \langle t, g\rangle = 0 \Leftrightarrow \sum_{i\in\mathcal{I}_+} \frac{n(i)}{N}\langle f_i, g\rangle = 0 \Leftrightarrow \langle f_i, g\rangle = 0\ \forall i \in I_+$. The face F_g may contain additional f_i's.

Thus $\mathbf{F}_t$ is the smallest face of $\mathbf{P}_\Delta$ containing $\{f_i, i \in I_+\}$, and so its facial set is $F_t = F_\Delta(I_+)$. Identifying $\mathbf{F}_t$ is therefore equivalent to identifying $F_\Delta(I_+)$.

Fact 2: If $\Delta' \subset \Delta$, then for any $S \subset I$, $F_\Delta(S) \subseteq F_{\Delta'}(S)$

Lemma 3.6. *Let Δ and Δ' be simplicial complexes on the same vertex set with $\Delta' \subseteq \Delta$, and denote by f_i, f_i' $(i \in I)$ the columns of the design matrices of the corresponding hierarchical models. Then there is a linear map $\phi : \mathbf{R}^h \to \mathbf{R}^{h'}$ with $\phi(f_i) = f_i'$. In fact, ϕ is a coordinate projection. In particular, the marginal polytope $\mathbf{P}_{\Delta'}$ is a coordinate projection of $\mathbf{P}_\Delta$. Thus, for any $S \subseteq I$, we have $F_\Delta(S) \subseteq F_{\Delta'}(S)$*

The lemma clearly follows from Lemma 1.7. This lemma says that we can find an outer approximation to $\mathbf{F}_t$ w.r.t Δ if $\Delta' \subset \Delta$ by looking at $F_{\Delta'}(I_+)$. In the other direction, we can find an inner approximation to $\mathbf{F}_t$ w.r.t Δ' by looking at $F_\Delta(I_+)$.

Fact 3: If Δ is reducible into several components, then $F_\Delta(I_+)$ can be obtained from the facial set relative to these components:

$$F_\Delta(T) = \pi_{V_1}^{-1}(F_{\Delta|_{V_1}}(T_1)) \cap \pi_{V_2}^{-1}(F_{\Delta|_{V_2}}(T_2)).$$

Let us now explain how this is obtained.

Definition 3.7. Let $V' \subset V$. The *restriction* or *induced sub-complex* is $\Delta|_{V'} = \{S \in \Delta \mid S \subseteq V'\}$. The sub-complex $\Delta|_{V'}$ is *complete*, if $\Delta|_{V'}$ contains V' (and thus all subsets of V'). For brevity, in this case we say that V' is *complete* in Δ.

Definition 3.8. A subset $S \subset V$ is a *separator* of Δ if there exist $V_1, V_2 \subset V$ with $V_1 \cap V_2 = S$, $\Delta = \Delta|_{V_1} \cup \Delta|_{V_2}$ and $V_1 \neq S \neq V_2$. A simplicial complex that has a complete separator is called *reducible*. By extension, we also call the hierarchical model reducible.

Definition 3.9. A hierarchical model is *decomposable* if Δ can be written as a union $\Delta = \Delta_1 \cup \Delta_2 \cup \ldots \Delta_r$ of induced sub-complexes $\Delta_i = \Delta|_{V_i}$ in such a way that

1. each Δ_i is a complete simplex: $\Delta_i = \{S \subseteq V_i\}$; and
2. $(\Delta_1 \cup \cdots \cup \Delta_i) \cap \Delta_{i+1}$ is a complete simplex.

In other words, Δ arises by iteratively gluing simplices along complete sub-simplices.

Faces of a reducible hierarchical model are combinations of the faces of its two parts:

Proposition 3.10. *Suppose that Δ has a complete separator S that separates V into V_1 and V_2. Each face of $\mathbf{P}_{\Delta|_{V_1}}$ corresponds to an inequality*

$$\sum_{j \in J_{\Delta|_{V_1}}} g_j^{(1)} t_j \geq c_1.$$

The same inequality also defines a face of $\mathbf{P}_\Delta$. Similarly, each face of $\mathbf{P}_{\Delta|_{V_2}}$ defines a face of $\mathbf{P}_\Delta$. Each face of $\mathbf{P}_\Delta$ either arises in this way, or it is the intersection of two such faces, one induced by $\mathbf{P}_{\Delta|_{V_1}}$ and one induced by $\mathbf{P}_{\Delta|_{V_2}}$.

Proof. See Eriksson et al. (2006), Lemma 8. □

In the sequel, for any $V' \subseteq V$ and $i \in I = \prod_{v \in V} I_v$, it will be convenient to use the seemingly more complicated notation $\pi_{V'}(i) = (i_v, v \in V')$ for the marginal cell $i_{V'} \in I_{V'} := \prod_{v \in V'} I_v$. Similarly, for a set $S \subseteq I$, the restriction to V' is $\pi_{V'}(S) := \{\pi_{V'}(i) : i \in S\}$. For $T \subset I_{V'}$, the opposite action yields $\pi_{V'}^{-1}(T) = \{i \in I \mid i_{V'} \in T\}$.

We next translate Proposition 3.10 to the language of facial sets:

Lemma 3.11. *Suppose that Δ has a complete separator S that separates V into V_1 and V_2.*

1. *If $F \subseteq I$ is facial with respect to Δ, then $\pi_{V_1}(F)$ and $\pi_{V_2}(F)$ are facial with respect to $\Delta|_{V_1}$ and $\Delta|_{V_2}$.*

2. *Conversely, if $F_1 \subseteq I_{V_1}$ and $F_2 \subseteq I_{V_2}$ are facial with respect to $\Delta|_{V_1}$ and $\Delta|_{V_2}$, then $\pi_{V_1}^{-1}(F_1) \cap \pi_{V_2}^{-1}(F_2)$ is facial with respect to Δ.*

Thus, for any $T \subseteq I$, let $T_1 = \pi_{V_1}(T)$ and $T_2 = \pi_{V_2}(T)$.

$$F_\Delta(T) = \pi_{V_1}^{-1}(F_{\Delta|_{V_1}}(T_1)) \cap \pi_{V_2}^{-1}(F_{\Delta|_{V_2}}(T_2)).$$

Proof. Consider an inequality as in Proposition 3.10 that defines a face $\mathbf{F}$ of $\mathbf{P}_\Delta$ as well as a face $\mathbf{F}_1$ of $\mathbf{P}_{\Delta_1}$. Then the corresponding facial sets F and F_1 satisfy $F = \pi_{V_1}^{-1}(F_1)$; because in order to check whether some f_i, $i \in I$, satisfies the inequality, we only need to look at the components involving V_1; that is, we only need to look at $\pi_{V_1}(i)$. □

Lemma 3.11 easily generalizes to more than one separator and thus to more than two components and it becomes particularly simple when these components are complete. Indeed, in that case, $F_{\Delta|_{V_1}}(T_1) = T_1$ and taking the preimage we obtain

$$\pi_{V_1}^{-1}(\pi_{V_1}(T)) = \{i \in I : \exists i' \in T \text{ such that } \pi_{V_1}(i) = \pi_{V_1}(i')\} \supseteq T.$$

The following lemma is an immediate consequence of Lemma 3.11.

Lemma 3.12. *Let Δ be a decomposable model with decomposition $\Delta = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_r$ where Δ_i is a complete simplex on V_i, and let $\pi_i = \pi_{V_i}$ be the corresponding marginalization map. Then, for any $T \subseteq I$,*

$$F_\Delta(T) = \pi_1^{-1}(\pi_1(T)) \cap \pi_2^{-1}(\pi_2(T)) \cap \cdots \cap \pi_r^{-1}(\pi_r(T)).$$

3.2.2 Finding an inner approximation to $\mathbf{F}_t$

The basic idea is to find a separator S in Δ, complete it and consider the augmented model $\Delta_S = \Delta \cup \{M : M \subseteq S\}$ in which S is a complete separator. We can apply Lemma 3.11 to find the facial set $F_{\Delta_S}(I_+)$, and this will be our inner approximation of $F_\Delta(I)$.

We refine this inner approximation by considering another separator S' and create the following chain of inner approximations:

$$\begin{gathered} G'_0 := I_+, \\ G_1 := F_{\Delta_S}(G'_0),\ G'_1 := F_{\Delta_{S'}}(G_1), \\ G_2 := F_{\Delta_S}(G'_1),\ G'_2 := F_{\Delta_{S'}}(G_2), \\ \vdots \end{gathered}$$

that satisfy

$$I_+ \subseteq G_1 \subseteq G'_1 \subseteq G_2 \subseteq \cdots \subseteq F_t,$$

where all inclusions except the last one are due to the definition of $F_{\Delta_S}(T)$ or $F_{\Delta_{S'}}(T)$ as the smallest facial sets containing T in Δ_S or $\Delta_{S'}$. The last inclusion is a consequence of Lemma 3.6 since both Δ_S and $\Delta_{S'}$ contain Δ. This chain of approximations has to stabilize at a certain point; that is, after a certain number of iterations, the approximations will not improve any more. The limit, which we denote by $F_{S,S'}(I^+) := \bigcup_i G_i = \bigcup_i G'_i$, can be characterized as the smallest subset of I that contains I^+ and is facial both with respect to Δ_S and $\Delta_{S'}$.

3.2.3 Finding an outer approximation to $\mathbf{F}_t$

The basic idea is to choose $\Delta' \subset \Delta$ so that according to Lemma 3.6, $F_\Delta(I_+) \subset F_{\Delta'}(I_+)$. In the smaller simplicial complex Δ', we can obtain the exact $F_{\Delta'}(I_+)$. Then we lift this facial set by using the following lemma.

Lemma 3.13. *Let $V' \subseteq V$. For $K \subset I$, we have*

$$F_{\Delta|_{V'}}(K) = \pi^{-1}_{V'}(F'_{\Delta|_{V'}}(\pi_{V'}(K))).$$

Here, $F'_{\Delta|_{V'}}$ denotes the facial set when $\Delta_{V'}$ is considered as a simplicial complex on V', and $F_{\Delta|_{V'}}$ denotes the facial set when $\Delta|_{V'}$ is considered as a simplicial complex on V.

Proof. Given a model Δ on V. Let $V' \subset V$. Consider the model $\Delta|_{V'}$. Let $\mathcal{A} = \{a_i, i \in I\}$ be the set of columns of the design matrix for the model $\Delta|_{V'}$ on V. Let $\mathcal{B} = \{a_{i'}, i' \in \pi_{V'}(I)\}$ be the set of columns of the design matrix for the model $\Delta|_{V'}$ on V'. For $K \subset I$, the two sets $\mathcal{A}_K = \{a_i, i \in K\}$ and $\mathcal{B}_K = \{a_{i'}, i' \in \pi_{V'}(K)\}$ are identical and therefore the smallest faces of the

marginal polytopes for $\Delta_{V'}$ on V or V' containing $\mathcal{A}_K$ and $\mathcal{B}_K$ respectively are the same.

By definition of $F'_{\Delta_{V'}}(\pi_{V'}(K))$, the smallest face containing $\mathcal{B}_K$ is defined by $\{a_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. By definition of $F_{\Delta_{V'}}(K)$, the smallest face containing $\mathcal{A}_K$ is $\{a_i, i \in F_{\Delta_{V'}}(K)\}$. Also because, for f_i a column of $\mathcal{A}$ and $f'_{i'}$ a column of $\mathcal{B}$, $f_i = f'_{i'} \Leftrightarrow i \in \pi^{-1}_{V'}(i')$, we have that $\{a_i,\ i \in \pi^{-1}_{V'}(F'_{\Delta_{V'}}(\pi_{V'}(K)))\} = \{b_{i'},\ i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. Therefore $F_{\Delta_{V'}}(K) = \pi^{-1}_{V'}(F'_{\Delta_{V'}}(\pi_{V'}(K)))$. □

Subsequently, since the outer approximation is not very accurate, one subset is not enough. So, we typically choose $V_1, \ldots, V_r \subseteq V$. Then $F_\Delta(I_+) \subseteq F_{\Delta|_{V_i}}(I_+)$ for $i = 1, \ldots, r$, and thus $F_\Delta(I_+) \subseteq \bigcap_{i=1}^r F_{\Delta|_{V_i}}(I_+) =: F_{V_1,\ldots,V_r;\Delta}(I_+)$.

3.2.4 Comparing the two approximations

Suppose that we have computed two approximations F_1, F_2 of F_t such that $F_1 \subseteq F_t \subseteq F_2$. If we are in the lucky case that $F_1 = F_2$, then we know that $F_t = F_1 = F_2$. In general, the cardinality of $F_2 \setminus F_1$ indicates the quality of our approximations.

F_1, F_2 and F_t can also be compared by the ranks of the matrices A_{F_1}, A_{F_2} and A_{F_t} obtained from A by keeping only the columns indexed by F_1, F_2 and F_t, respectively. Clearly, $\operatorname{rank} A_{F_1} \leq \operatorname{rank} A_{F_t} \leq \operatorname{rank} A_{F_2}$. Note that $\operatorname{rank} A_{F_2}$ equals the dimension of the corresponding face $\mathbf{F}_2$ of $\mathbf{P}$, and $\operatorname{rank} A_{F_t}$ equals the dimension of $\mathbf{F}_t$. But F_1 does not necessarily correspond to a face of $\mathbf{P}$. Nevertheless, we can bound the codimension of $\mathbf{F}_t$ in $\mathbf{F}_2$ by

$$\dim \mathbf{F}_2 - \dim \mathbf{F}_t \leq \operatorname{rank} A_{F_2} - \operatorname{rank} A_{F_1}.$$

In particular, if $\operatorname{rank} A_{F_2} = \operatorname{rank} A_{F_1}$, then we know that $F_t = F_2$. In this case, our approximations give us a precise answer, even if $F_1 \neq F_2$ and the lower approximation F_1 is not tight.

4 How to compute the extended mle? An Example

Consider two binary random variables, and let $\Delta = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. The hierarchical model $\mathcal{E}_\Delta$ is the *saturated model*; that is, it contains all possible probability distributions with full support. Then

$$\tilde{A} = \begin{array}{c} \overbrace{}^{f_{00}} \overbrace{}^{f_{01}} \overbrace{}^{f_{10}} \overbrace{}^{f_{11}} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \theta_{00} \\ \theta_{01} \\ \theta_{10} \\ \theta_{11} \end{array}$$

The marginal polytope is a 3-simplex (a tetrahedron) with facets

$$\mathbf{F}_{00} : 1 - t_{01} - t_{10} + t_{11} \geq 0, \quad \mathbf{F}_{01} : t_{01} - t_{11} \geq 0,$$
$$\mathbf{F}_{10} : t_{10} - t_{11} \geq 0, \quad \mathbf{F}_{11} : t_{11} \geq 0.$$

Each of the corresponding facets contains three columns of $\tilde{A}$. In fact, the facet $\mathbf{F}_i$ in the above list does not contain the column f_i of A.

The EMLE of the saturated model is just the empirical distribution; that is, $p_* = \frac{1}{N}n$. Suppose that t lies on the facet $\mathbf{F}_{00}$ (i.e. $n = (0, n_{01}, n_{10}, n_{11})$ with $n(01), n(10), n(11) > 0$). If $p_{\theta^{(s)}} \to p_*$, then $p_{\theta^{(s)}}(00) \to 0$, while all other probabilities converge to a non-zero value. It follows that

$$\begin{aligned}
\theta_{00}^{(s)} &= \log p_{\theta^{(s)}}(00) \to -\infty, \\
\theta_{01}^{(s)} &= \log \frac{p_{\theta^{(s)}}(01)}{p_{\theta^{(s)}}(00)} \to +\infty, \\
\theta_{10}^{(s)} &= \log \frac{p_{\theta^{(s)}}(10)}{p_{\theta^{(s)}}(00)} \to +\infty, \\
\theta_{11}^{(s)} &= \log \frac{p_{\theta^{(s)}}(11)p_{\theta^{(s)}}(00)}{p_{\theta^{(s)}}(01)p_{\theta^{(s)}}(10)} \to -\infty.
\end{aligned}$$

On the other hand, $\theta_{01}^{(s)} + \theta_{00}^{(s)} = \log p_{\theta^{(s)}}(01)$ converges to a finite value, as do $\theta_{10}^{(s)} + \theta_{00}^{(s)} = \log p_{\theta^{(s)}}(10)$ and $\theta_{11}^{(s)} + \theta_{01}^{(s)} = \log p_{\theta^{(s)}}(11)/p_{\theta^{(s)}}(10)$.

Proceeding similarly for the other facets, one can show for the limits $\theta_{ij} := \lim_{s\to\infty} \theta_{ij}^{(s)}$:

	θ_{00}	θ_{01}	θ_{10}	θ_{11}	finite parameter combinations:
$\mathbf{F}_{00}$	$-\infty$	$+\infty$	$+\infty$	$-\infty$	$\theta_{01}^{(s)} + \theta_{00}^{(s)}$, $\theta_{10}^{(s)} + \theta_{00}^{(s)}$, $\theta_{11}^{(s)} + \theta_{01}^{(s)}$
$\mathbf{F}_{01}$	finite	$-\infty$	finite	$+\infty$	$\theta_{00}^{(s)}$, $\theta_{10}^{(s)}$, $\theta_{01}^{(s)} + \theta_{11}^{(s)}$
$\mathbf{F}_{10}$	finite	finite	$-\infty$	$+\infty$	$\theta_{00}^{(s)}$, $\theta_{01}^{(s)}$, $\theta_{10}^{(s)} + \theta_{11}^{(s)}$
$\mathbf{F}_{11}$	finite	finite	finite	$-\infty$	$\theta_{00}^{(s)}$, $\theta_{10}^{(s)}$, $\theta_{01}^{(s)}$

Each line of the last column contains three combinations of the parameters $\theta_i^{(s)}$ that converge to a finite value. Any other parameter combination that

converges is a linear combination of these three. This can be seen by using the coordinates $\mu_i = \log \frac{p(i)}{p(0)}$ and applying the following lemma.

Lemma 4.1. *Suppose that $\theta^{(s)}$, $s \in \mathbb{N}$, are parameter values such that $p_{\theta^{(s)}} \to p^*$ as $s \to \infty$. For any $i \in L_t$, the linear combination*

$$\mu_i^{(s)} = \langle \theta^{(s)}, f_i \rangle$$

has a well-defined finite limit as $s \to \infty$. Any linear combination of the $\theta_i^{(s)}$ that has a well-defined finite limit (that is, a limit that is independent of the choice of the sequence $\theta^{(s)}$) is itself a linear-combination of the $\mu_i^{(s)}$ with $i \in L_t$.

For example, on the facet $\mathbf{F}_{01}$, consider the parameters

$$\mu_{10} = \log p(10)/p(00) = \theta_{10}, \qquad \mu_{11} = \log p(11)/p(00) = \theta_{10} + \theta_{01} + \theta_{11},$$
$$\mu_{01} = \log p(01)/p(00) = \theta_{01}.$$

Then μ_{10} and μ_{11} are identifiable parameters on $\mathcal{E}_{F_{01}}$, and μ_{01} diverges close to $\mathbf{F}_{01}$. By Lemma 4.1, the linear combinations that are well-defined are $\mu_{10} = \theta_{10}$ and $\mu_{11} = \theta_{10} + (\theta_{01} + \theta_{11})$. The above table also lists θ_{00}, which is not a linear combination of those but that is fine because it is not free.

We obtain similar results for the facets $\mathbf{F}_{01}$ and $\mathbf{F}_{11}$. The results are summarized in the following table:

facet	μ_{01}	μ_{10}	μ_{11}
$\mathbf{F}_{01}$	$-\infty$	finite	finite
$\mathbf{F}_{10}$	finite	$-\infty$	finite
$\mathbf{F}_{11}$	finite	finite	$-\infty$

Of course, by definition of the μ_is, we cannot consider the facet $\mathbf{F}_{00}$ where $n(00) = 0$. To study $\mathbf{F}_{00}$, we have to choose another zero cell and redefine the parameters μ_i.

The situation is more complicated for faces smaller than facets, because sending a single parameter to plus or minus infinity can be enough to send the distribution to a face F of higher codimension, as we will see below. The remaining parameters then determine the position within $\mathcal{E}_{\Delta,F}$. Thus, in this case there are more remaining parameters than the dimension of $\mathcal{E}_{\Delta,F}$.

For example, the data vector $n = (n_{00}, 0, n_{10}, 0)$ (with $n_{00}, n_{10} > 0$) lies on the face $\mathbf{F} = \mathbf{F}_{01} \cap \mathbf{F}_{11}$ of codimension two. If $p_{\theta^{(s)}} \to p_*$, then

$$\begin{aligned}
\theta_{00}^{(s)} &= \log p_{\theta^{(s)}}(00) \to \log \frac{n_{00}}{N}, \\
\theta_{01}^{(s)} &= \log \frac{p_{\theta^{(s)}}(01)}{p_{\theta^{(s)}}(00)} \to -\infty, \\
\theta_{10}^{(s)} &= \log \frac{p_{\theta^{(s)}}(10)}{p_{\theta^{(s)}}(00)} \to \log \frac{n_{10}}{n_{00}}.
\end{aligned}$$

However, the limit of $\theta_{11}^{(s)} = \log \frac{p_{\theta^{(s)}}(11) p_{\theta^{(s)}}(00)}{p_{\theta^{(s)}}(01) p_{\theta^{(s)}}(10)}$ is not determined. The only constraint is that $\theta_{11}^{(s)}$ cannot go to $+\infty$ faster than $\theta_{01}^{(s)}$ goes to $-\infty$, since $p_{\theta_{11}^{(s)}} = \exp(\theta_{00}^{(s)} + \theta_{01}^{(s)} + \theta_{10}^{(s)} + \theta_{11}^{(s)})$ has to converge to zero.

With the same data vector $n = (n_{00}, 0, n_{10}, 0)$, suppose we use a numerical algorithm to optimize the likelihood function by optimizing the parameters θ_j in turn. To be precise, we order the parameters θ_j in some way. For simplicity, say that the parameters are $\theta_1, \theta_2, \ldots, \theta_h$. Then we let

$$\theta_j^{(k+1)} = \arg\max_{y \in \mathbf{R}} l(\theta_1^{(k+1)}, \ldots, \theta_{j-1}^{(k+1)}, y, \theta_{j+1}^{(k)}, \ldots, \theta_h^{(k)})$$

(this is called the *non-linear Gauss-Seidel method*). Let us choose the ordering $\theta_{01}, \theta_{10}, \theta_{11}$ (note that $\theta_{00} = -k(\theta)$ is not a free parameter). We start at $\theta_{01}^{(0)} = \theta_{10}^{(0)} = \theta_{11}^{(0)} = 0$. In the first step, we only look at θ_{01}. That is, we want to solve

$$\begin{aligned}
0 = \frac{\partial}{\partial \theta_{01}} l(\theta) &= -\frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(0)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})} \\
&= -\frac{2\exp(\theta_{01}^{(1)})}{1 + 2\exp(\theta_{01}^{(1)})}. \quad (22)
\end{aligned}$$

Clearly, the derivative is negative for any finite value of $\theta_{01}^{(1)}$, and thus the critical equation has no finite solution. If we try to solve this equation numerically, we will find that $\theta_{01}^{(1)}$ will be a large negative number. Next, we

look at θ_{10}. We fix the other variables and try to solve

$$0 = \frac{\partial}{\partial \theta_{10}} l(\theta) = \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(0)})}$$
$$\approx \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)})}{1 + \exp(\theta_{10}^{(1)})},$$

where we have used that $\theta_{01}^{(1)}$ is a large negative number. This equation always has a unique solution

$$\theta_{10}^{(1)} \approx \log \frac{n_{10}}{N - n_{10}}.$$

Finally, we look at θ_{11}. We have to solve

$$0 = \frac{\partial}{\partial \theta_{11}} l(\theta) = -\frac{\exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})} \approx 0.$$

Actually, this equation again has no solution, and the numerical solution for $\theta_{11}^{(1)}$ should be close to numerical minus infinity. However, since $\theta_{01}^{(1)}$ is already close to $-\infty$, the equation is already approximately satisfied. Thus, there is no need to change θ_{11}. In simulations, we observed that usually $\theta_{11}^{(1)}$ will be negative, but not as negative as $\theta_{01}^{(1)}$. In theory, we would have to iterate and now optimize θ_{01} again. But the values will not change much, since the critical equations are already satisfied to a high numerical precision after one iteration.

It is not difficult to see that the result is different if we change the order of the variables. If θ_{11} is optimized before θ_{01}, then θ_{11}^{1} will in any case be a large negative number.

For general data, the derivative of with respect to θ_{01} (equation (22)) takes the form

$$\frac{\partial}{\partial \theta_{01}} l(\theta) = \frac{t_{01}}{N} - \frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(0)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}.$$

Setting this derivative to zero and solving for $\theta_{01}^{(1)}$ leads to a linear equation in $\theta_{01}^{(1)}$ with symbolic solution

$$\theta_{01}^{(1)} = \log \frac{1 + \exp(\theta_{10}^{(0)})}{1 + \exp(\theta_{10}^{(0)} + \theta_{11}^{(0)})} \frac{\frac{t_{01}}{N}}{1 - \frac{t_{01}}{N}}.$$

In fact, for any hierarchical model, the likelihood equation is linear in any single parameter θ_j, as long as all other parameters are kept fixed (more generally this is true when the design matrix A is a 0-1-matrix). Instead of optimizing the likelihood numerically with respect to one parameter, it is possible to use these symbolic solutions. This leads to the Iterative Proportional Fitting Procedure (IPFP). In our example, the IPFP would lead to a division by zero right in the first step, indicating that the MLE does not exist.

A Some proofs

A.1 Proof of Theorem 1.10

Theorem 1.10 goes back to Barndorff:exponential families, (1978), who studies the closure of much more general exponential families. The case of a discrete exponential family is much easier.

The theorem follows from the following lemmas:

Lemma A.1. *Let* $p \in \overline{\mathcal{E}_A}$. *Then* $p \in \mathcal{E}_{A,\mathrm{supp}(p)}$.

Lemma A.2. *Let* $p \in \overline{\mathcal{E}_A}$. *Then* $\mathcal{E}_{A,\mathrm{supp}(p)} \subseteq \overline{\mathcal{E}_A}$.

Lemma A.3. *Let* $p \in \overline{\mathcal{E}_A}$. *Then* $\mathrm{supp}(p)$ *is facial.*

Lemma A.4. *If* F *is facial, then there exists* $p \in \overline{\mathcal{E}_A}$ *with* $\mathrm{supp}(p) = F$.

Indeed, Lemma A.1 shows that $\overline{\mathcal{E}_A} \subseteq \bigcup_F \mathcal{E}_{A,F}$, where the union is over all support sets F. Lemma A.2 shows the converse containment, so that $\overline{\mathcal{E}_A} = \bigcup_F \mathcal{E}_{A,F}$. It remains to see that a subset $F \subseteq I$ is a support set if and only if F is facial. This follows from Lemmas A.3 and A.4.

In the proofs of Lemmas A.1 to A.4, we need the following easy lemma which follows immediately from the fact that $p \in \mathcal{E}_A$ if and only if $\log p$ belongs to the span of the columns of A and therefore if and only if $\log p$ is perpendicular to the $ker(A)$.

Lemma A.5. $p \in \mathcal{E}_A$ *if and only if* $\log(p) \perp \ker A$.

Proof of Lemma A.1. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, and let $F = \mathrm{supp}(p)$. Then $\mathcal{E}_{A,F}$ is the exponential family $\mathcal{E}_{A_F}$, where A_F consists of

the columns of A indexed by F. Any $v \in \ker A_F$ can be extended by zeros to $v' \in \ker A$. By Lemma A.5,

$$0 = \langle \log(p_k), v' \rangle = \sum_{i \in F} \log(p_k(i)) v(i) \to \langle \log(p), v \rangle.$$

Thus, $\log(p) \perp \ker A_F$, which implies $p \in \mathcal{E}_{A,F}$. □

Proof of Lemma A.2. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, let $F = \operatorname{supp}(p)$, and let $q \in \mathcal{E}_{A,F}$. Then there exists parameters θ with $\log(q(i)) - \log(p(i)) = \langle \theta, f_i \rangle$ for all $i \in F$. For any k, there exists a positive constant c_k such that $q_k := c_k p_k \exp(\langle \theta, A \rangle) \in \mathcal{E}_A$. Then $q_k \to q$ as $k \to \infty$, and so $q \in \overline{\mathcal{E}_A}$. □

Proof of Lemma A.3. Let $p = \lim_{k\to\infty} p_k$, where $p_k \in \mathcal{E}_A$, and let $F = F_A(\operatorname{supp}(p))$. Then $x = \frac{1}{|\operatorname{supp}(p)|} \sum_{i \in \operatorname{supp}(p)} f_i$ is an interior point of the face corresponding to F, and thus there exist positive coefficients $\lambda_i > 0$, $i \in F$, with $x = \sum_{i \in F} \lambda_i f_i$. The vector $v = (v_i, i \in I)$ defined by

$$v_i = \begin{cases} \frac{1}{|\operatorname{supp}(p)|} - \lambda_i, & i \in \operatorname{supp}(p), \\ -\lambda_i, & i \in F \setminus \operatorname{supp}(p), \\ 0, & i \notin F, \end{cases}$$

satisfies $Av = x - x = 0$. By Lemma A.5, $\log(p_k) \perp v$ for all k. In particular,

$$\sum_{i \in F \setminus \operatorname{supp}(p)} \lambda_i \log(p_k(i)) = \sum_{i \in \operatorname{supp}(p)} \log(p_k(i)) v_i \to \sum_{i \in \operatorname{supp}(p)} \log(p(i)) v_i.$$

On the other hand, note that each coefficient λ_i for $i \in F \setminus \operatorname{supp}(p)$ on the left hand side is positive, while $\log(p_k(i)) \to -\infty$ for $i \notin \operatorname{supp}(p)$. This shows that $F \setminus \operatorname{supp}(p) = \emptyset$. □

Proof of Lemma A.4. If F is facial, there exist $g \in \mathbf{R}^h$ and $c \in \mathbf{R}$ with $\langle g, f_i \rangle \geq c$ for all $i \in I$ and $\langle g, f_i \rangle = c$ if and only if $i \in F$. Let $\theta^{(s)} = -s \cdot g$. Then

$$k_F(\theta_{(s)}) + sc = \log \sum_{i \in I} \exp(-s\langle g, f_i \rangle + sc) \to \log |F|,$$

and so

$$\log p_{\theta^{(s)}}(i) = -s\langle g, f_i \rangle - k_F(\theta_{(s)}) = (sc - s\langle g, f_i \rangle) - (k_F(\theta_{(s)}) + sc) \to \begin{cases} -\log |F|, & \text{if } i \in F, \\ -\infty, & \text{if } i \notin F, \end{cases}$$

as $s \to \infty$. Thus, $p_{\theta^{(s)}}$ converges to the uniform distribution on F. □

A.2 Proof of Theorem 1.11

By definition, any EMLE p_* belongs to the closure of the model. According to Theorem 1.10, the support of p_* is facial. If $\mathrm{supp}(p)$ does not contain $\mathrm{supp}(n)$, then the log-likelihood goes to minus infinity, $\tilde{l}(p) = -\infty$, and so p does not maximize the likelihood, Therefore, $\mathrm{supp}(p_*)$ is a facial set containing $I_+ = \mathrm{supp}(n)$. Thus, $F_t \subseteq \mathrm{supp}(p_*)$.

By Lemma A.1, p_* belongs to $\mathcal{E}_{\Delta,\mathrm{supp}(p_*)}$, which is parametrized by a vector θ, see Theorem 1.10. On $\mathcal{E}_{\Delta,\mathrm{supp}(p_*)}$, the log-likelihood function in terms of this parameter θ is

$$l_F(\theta) = \sum_{j\in J} \theta_j t_j - N k_F(\theta).$$

l_F is strictly concave, and so it has a unique maximum. The critical equations are

$$t_j = N\frac{\partial k_F(\theta)}{\partial \theta_j} = N p^*_J = N A p^* \quad \text{and thus} \quad A p_* = \frac{t}{N},$$

where p^*_J is the vector of J-marginal probabilities, proving the first property. Note that these equations are independent of the parameters and the support of p_*. We now show that any solution to these equations is supported on the same face of $\mathbf{P}$ as $\frac{t}{N}$.

Let p be a probability distribution on I such that $\mathrm{supp}(p)$ does not contain F_t. This means that there is a linear inequality $\langle g, t\rangle \geq c$ that is valid on $\mathbf{P}$ and such that

- $\langle g, f_i\rangle = c$ for all $i \in F_t$;
- $\langle g, f_i\rangle > c$ for some $i \in \mathrm{supp}(p)$.

Then

$$\langle g, Ap\rangle = \sum_i \langle g, f_i\rangle p(i) > c = \frac{1}{N}\sum_i n(i)\langle g, f_i\rangle = \langle g, \frac{t}{N}\rangle,$$

which implies $Ap \neq \frac{t}{N}$. This shows $\mathrm{supp}(p_*) \subseteq F_t$ and finishes the proof of $\mathrm{supp}(p_*) = F_t$.

We have now shown the two properties, and it remains to argue that the EMLE is unique. But this follows from the fact that $\mathrm{supp}(p_*)$ is equal to F_t, and l_F is strictly concave, such that the likelihood has a unique maximizer on $\mathcal{E}_{\Delta,F_t}$.

B References

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research, 9:485-516, 2008.

O.E. Barndorff-Nielsen. Information and Exponential Families. Wiley, Chichester, first edition, 1978.

I. Csiszar and F. Matus. Closures of exponential families. Annals of Probability, 33: 582-600, 2005.

A. Dobra, E.A. Erosheva, and S.E. Fienberg. Disclosure limitation methods based on bounds for large contingency tables with application to disability data. In Proceedings of conference on the new frontiers of statistical data mining, pages 93?116, 2003.

A. Dobra and A. Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. Ann. Appl. Stat., 5(2A):969-993, 06 2011.

N. Eriksson, S. Fienberg, A. Rinaldo, and S. Sullivant. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. J. Symbolic Comput., 41: 222-233, 2006.

S. E. Fienberg and A. Rinaldo. Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. J. of Statistical PLanning and Inference, 137:3430-3445, 2007.

S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. Annals of Statistics, 40:9961023, 2012.

C. J. Geyer. Likelihood inference in exponential families and directions of recession. Electron. J. Statist., 3:259-289, 2009.

S. J. Haberman. The Analysis of Frequency Data. Univ. Chicago Press, Chicago, IL, 1974. S.L. Lauritzen. Graphical Models. Oxford Science Publications, first edition, 1996.

G. Letac and H. Massam. Bayes regularization and the geometry of discrete hierarchical loglinear models. Annals of Statistics, 40:861-890, 2012.

H. Massam and N. Wang. A local approach to estimation in discrete

loglinear models. Preprint, 2015. arXiv:1504.05434.

J. Rauh, T. Kahle, and N. Ay. Support sets of exponential families and oriented matroids. International Journal of Approximate Reasoning, 52(5):613-626, 2011.

P. Ravikumar, M. J. Wainwright, J. D. Lafferty, et al. High-dimensional ising model selection using l1-regularized logistic regression. The Annals of Statistics, 38(3):1287- 1319, 2010.

G. Ziegler. Lectures on Polytopes. Springer, second edition, 1998.