The existence of the maximum likelihood estimate in hierarchical loglinear models is crucial to the reliability of inference for this model. Determining whether the estimate exists is equivalent to finding whether the sufficient statistics vector $t$ belongs to the boundary of the marginal polytope of the model. The dimension of the smallest face $F_t$ containing $t$ determines the dimension of the reduced model which should be considered for correct inference. For higher-dimensional problems, it is not possible to compute $F_t$ exactly. Massam and Wang (2015) found an outer approximation to $F_t$ using a collection of sub-models of the original model. This paper refines the methodology to find an outer approximation and devises a new methodology to find an inner approximation. The inner approximation is given not in terms of a face of the marginal polytope, but in terms of a subset of the vertices of $F_t$.

Knowing $F_t$ exactly indicates which cell probabilities have maximum likelihood estimates equal to 0. When $F_t$ cannot be obtained exactly, we can use, first, the outer approximation $F_2$ to reduce the dimension of the problem and, then, the inner approximation $F_1$ to obtain correct estimates of cell probabilities corresponding to elements of $F_1$ and improve the estimates of the remaining probabilities corresponding to elements in $F_2 \setminus F_1$. Using both real-world and simulated data, we illustrate our results, and show that our methodology scales to high dimensions.

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

*H. Massam gratefully acknowledges support from an NSERC Discovery Grant*
1. Introduction. Discrete hierarchical models are an essential tool for the analysis of categorical data given under the form of a contingency table. The study of these models goes back more than a century, and a detailed history of their development is given in Fienberg and Rinaldo (2007). Nowadays, discrete hierarchical models are used for the analysis of large sparse contingency tables where many, if not most, of the entries are small or zero counts. It is well-known that in such cases, the maximum likelihood estimate (henceforth abbreviated MLE) of the parameters may not exist. The non existence of the MLE has problematic consequences for inference, clearly for estimation, but also for testing and model selection. Fienberg and Rinaldo (2012) list the statistical implications of the non existence of the MLE, such as the unreliability of the estimates of some of the parameters or the usage of the wrong degrees of freedom for testing one model against another. Geyer (2009) describes the problems attached to the nonexistence of the MLE and presents an R program that yields meaningful confidence intervals and tests. Letac and Massam (2012) study the statistical implications of the nonexistence of the MLE on model selection in Bayesian inference.

Fienberg and Rinaldo (2012) also give necessary and sufficient conditions for the existence of the MLE, which are not restricted to hierarchical models, but apply to all discrete exponential families (loglinear models). These conditions are extensions of results given earlier by Haberman (1974), Barndorff-Nielsen (1978) and Eriksson et al. (2006). They are essentially as follows. Denote by $I$ the set of outcomes of a statistical experiment, that is the set of cells of the contingency table where the data is classified. Let $I^+ = \{i_1, \ldots, i_N\}$ be the outcome of $N$ independent repetitions of the experiment within either a multinomial or Poisson setting. The data $I^+$ is summarized by the vector $t$ of sufficient statistics, which is of the form $t = \sum_{j=1}^N f_{ij}$ for some vectors $f_{ij}, i \in I$, determined by the given hierarchical loglinear model. Under these assumptions, the distribution of the data belongs to a natural exponential family with density

$$f(i_1, \ldots, i_N; \theta) = \exp\{\langle \theta, t \rangle - Nk(\theta)\}$$

with respect to the counting measure, where $\theta$ is a loglinear parameter. To each exponential family is associated a polytope $P$, called the marginal polytope, which is the convex hull of the vectors $f_{ii}, i \in I$. Furthermore, $P$ contains all possible realizations of $\frac{t}{N}$ for arbitrary repetitions of the statistical experiment. For given data and a given hierarchical model, the MLE then exists if and only if $\frac{t}{N}$ belongs to the relative interior of $P$. If the MLE does not exist, then $\frac{t}{N}$ belongs to the relative interior of a face denoted $F_t$. It is the smallest face of $P$ containing $\frac{t}{N}$, and it is proper
(i.e. $F_t \neq P$). Thus, determining whether, for a given data set, the MLE of the parameter of a discrete hierarchical loglinear model exists is equivalent to determining whether $\frac{t}{N}$ belongs to a proper face of $P$. The parameter may be the loglinear parameter $\theta$, or the cell probabilities $p = (p(i), i \in I)$ obeying the constraints of the model and in 1-1 correspondence with $\theta$. The MLE can thus be thought of in terms of $\theta$, or in terms of $p$.

If the MLE does not exist, it is still possible to compute the extended MLE (EMLE) (Barndorff-Nielsen, 1978; Lauritzen, 1996; Csiszár and Matúš, 2008), which is a probability distribution that maximizes the likelihood over the closure of the hierarchical model (that is, the EMLE can be approximated arbitrarily well by distributions from the hierarchical model). The support of the EMLE is given by the set $F_t = \{i \in I : f_i \in F_t\}$, called the facial set of $F_t$. When this support is known, computing the EMLE is equivalent to an ordinary MLE computation on a smaller exponential family $\mathcal{E}_{F_t}$, of dimension $\dim(F_t)$, generated by a measure with support $F_t$ (Geyer, 2009). Therefore, precise knowledge of $F_t$ and $F_t$ yields which is the proper dimension of the model to be used in testing and which outcomes $i \in I$ are attributed a probability of 0 by the EMLE, and it allows us to compute the EMLE. One should also note that the usual regularity conditions used for the asymptotic properties of the MLE, which are not satisfied for the given model when the MLE does not exist, are satisfied for the reduced model $\mathcal{E}_{F_t}$.

The problem is then to find $F_t$. This is easy when the face lattice of $P$ is known or can be computed using a standard discrete geometry toolkit such as, for example, polymake (Gawrilow and Joswig, 2000). For some classes of marginal polytopes, the face lattice is known, for example for decomposable models and no-three-way-interaction models with small variables (Vlach, 1986). For binary variables, the marginal polytope is a cut polytope (Deza and Laurent, 2009). Other authors have studied convex support polytopes, which replace marginal polytopes for more general exponential families. Notably, many such polytopes have been described for exponential random graph models, see, for example, Karwa and Slavkovic (2016) and papers cited therein. When the face lattice of $P$ cannot be computed, algorithms to compute $F_t$ that are based on linear programming have been proposed by Eriksson et al. (2006), by Geyer (2009), and by Fienberg and Rinaldo (2012). These methods, however, become computationally infeasible in large dimensions, which happens, in our experience, for hierarchical models when the set of random variables $V$ contains more than 16 binary variables (or correspondingly fewer larger variables).

For larger models, Massam and Wang (2015) propose to approximate $F_t$
by relating it to faces of smaller hierarchical models as follows. A hierarchical model for the discrete random variable $X = (X_v, v \in V)$ is determined by a set of interactions among its components $X_v, v \in V$, that is represented by a simplicial complex $\Delta$. Massam and Wang (2015) consider subsets $V_i, i = 1, \ldots, k,$ of $V$ containing less than 16 variables and the hierarchical models Markov with respect to the induced simplicial subcomplexes. Linear programming can be used to compute the smallest faces $F_{t_i}$ containing the corresponding sufficient statistic $t_i, i = 1, \ldots, k$. These faces, which a priori are faces of the marginal polytopes of the submodels, naturally correspond to faces of the original marginal polytope. Massam and Wang (2015) prove that the intersection of these is a face $F_2$ of $P$ containing $F_{t_i}$. Thus, if $F_2$ is a proper face of $P$, then $F_{t_i}$ is necessarily a proper face, and therefore the MLE does not exist. While Massam and Wang (2015) work with graphical models, we show that their results are also true for hierarchical loglinear models.

We call $F_2$ an outer approximation to $F_{t_i}$. This is similar to the notion of an outer approximation in optimization, which describes a polytope that contains the original polytope of interest. While the outer approximation polytope in optimization usually has the same dimension as the original polytope, the outer approximation face $F_2$ does not necessarily have the same dimension as $F_{t_i}$.

The purpose of this paper is to add to this outer approximation $F_2$ an inner approximation $F_1$ that is a subset of $F_{t_i}$. While $F_2$ is derived from looking at simplicial subcomplexes of $\Delta$, the inner approximation is constructed by enlarging the simplicial complex through added interactions. In particular, we propose a process of “completing a separator,” which leads to a decomposable simplicial complex which, in turn, can be studied by looking at the sub-simplices corresponding to its components with a small number of vertices in $V$. Thus, both $F_1$ and $F_2$ can be obtained by computing facial sets on smaller hierarchical models involving fewer nodes.

The inner and outer approximations to $F_{t_i}$ satisfy

$$F_1 \subset F_{t_i} \subset F_2.$$ 

Clearly, we want $F_1$ as large as possible and $F_2$ as small as possible to have as much information about $F_{t_i}$ as possible. In our simulations, we observe that $F_{t_i} = F_2$ most of the time and that $F_{t_i} = F_1$ quite often. The approximations $F_1$ and $F_2$ allow to bound the dimension of $F_{t_i}$, and thus knowledge from $F_1$ and $F_2$ can be taken into account whenever the dimension of $F_{t_i}$ plays a role, for example in hypothesis testing.

When the MLE does not exist, even though the maximum likelihood
procedure cannot be used to obtain a point estimate for the parameter vector $\theta$, some of its components $\theta_j$ may still be finite and well-defined in this situation. In Section 4 we introduce a loglinear parametrization $\mu$, different from $\theta$, that allows to say precisely which parameter combinations have a finite well-defined limit and thus remain meaningful for statistical inference. Moreover, we demonstrate that even when $F_t$ is unknown, the parametrization $\mu$ can be adjusted to incorporate knowledge that is available in the form of inner and outer approximations $F_1$ and $F_2$.

We extend the work of Fienberg and Rinaldo (2012) and that of Geyer (2009) in several directions: first, we construct approximations to $F_t$ in high dimensions when a direct computation of $F_t$ is not feasible. Second, we explicitly identify all parameter combinations that remain finite and meaningful when the MLE does not exist and $F_t$ is known, and we also discuss what can be said when only approximations to $F_t$ are available.

The remainder of this paper is organized as follows. In Section 2, we give preliminaries on hierarchical models, and faces and facial sets. Section 3 contains the original methodology to obtain the approximations $F_1$ and $F_2$. In Section 4, we show how to use $F_1$ and $F_2$ to identify the parameters of the hierarchical models that can be estimated and those that cannot. In Section 5, we present two examples. A simulated data set is used to assess how often our approximations succeed to identify the true facial set $F_t$. The NLTCS data set, studied by Dobra et al. (2003) and Dobra and Lenkoski (2011), illustrates how the outer approximation $F_2$ improves estimates of cell probabilities and log-linear parameters. Both of these examples have 16 nodes. In Section 6, we discuss how to apply the methodology to larger models and how to use for inference the information that it yields. Two examples illustrate this: simulated data from the graphical model of the $5 \times 10$ grid, and the real-world data set of voting records in the US Senate.

Appendix A describes the concrete parametrization that we use in the examples. Appendix B discusses the case of two binary variables to illustrate what happens to the usual parameters when the MLE does not exist. Appendix C discusses how to further improve the parametrization $\mu_L$ introduced in Section 2. Appendices D and E give further results for the examples from Section 5. Appendix F briefly summarizes the linear programming algorithm to compute $F_t$ by Fienberg and Rinaldo (2012). Appendix G contains proofs of the results of Barndorff-Nielsen (1978) that describe the closure of an exponential family and the EMLE.

Our results apply not only to hierarchical models, but to arbitrary discrete exponential families. In this paper, the focus is on hierarchical and graphical
models, which are the main application, and for which the construction of the inner and outer approximations can be described in terms of the underlying simplicial complex or graph.

2. Preliminaries. In the following four subsections, we recall basic facts about hierarchical models, discrete exponential families, polytopes and the closure of exponential families, and we define the extended MLE.

2.1. Hierarchical models and discrete exponential families. For details and proofs on the material in this subsection, we refer to Letac and Massam (2012) and Rauh et al. (2011). Let \( X = (X_v, v \in V) \) be a discrete random vector with components indexed by a finite set \( V = \{1, \ldots, p \} \). 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 a \( p \)-dimensional contingency table. For any \( D \subseteq V \), the subvector \( X_D = (X_v)_{v \in D} \) takes its values in \( I_D = \prod_{v \in D} I_v \). The \( D \)-marginal cell of \( i \in I \) will be denoted by \( i_D = (i_v)_{v \in D} \). The corresponding restriction is the coordinate projection map \( i \mapsto i_D \) and is denoted by \( \pi_D \).

Let \( \Delta \) be a simplicial complex on \( V \), that is, \( \Delta \) is a set of subsets \( D \subseteq V \) such that \( D \in \Delta \) and \( D' \subseteq D \) implies \( D' \in \Delta \). The joint distribution of \( X \) is hierarchical with underlying simplicial complex \( \Delta \) (or generating set \( \Delta \)) if the probability \( p(i) = P(X = i) \) of a single cell \( i = (i_v, v \in V) \) is of the form

\[
\log p(i) = \sum_{D \in \Delta} \theta_D(i_D)
\]

where \( \theta_D(i_D) \) is a function of the marginal cell \( i_D = (i_v, v \in D) \) only. To make precise the dependence on \( \theta \), we sometimes 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 \( \Delta \).

Equation (2) is essentially a linear condition on \( \log p(i) \). It is possible to parametrize the hierarchical model using a finite vector of parameters \( (\theta_j)_{j \in J} \) such that

\[
\log p_\theta(i) = \sum_{j \in J} \theta_j a_{j,i} - k(\theta),
\]

where \( A_\Delta = (a_{j,i})_{j \in J, i \in I} \) is a fixed real matrix (depending only on \( \Delta \)) and where \( k(\theta) = \log \sum_i \exp(\sum_j \theta_j a_{j,i}) \) ensures the normalization \( \sum_{i \in I} p_\theta(i) = 1 \). This parametrization is not unique. In the examples, we use an explicit parametrization that is used, for example, by Letac and Massam (2012). For convenience, we recall this parametrization in Appendix A.
An important subclass of hierarchical models 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 for any \( i, j \in V, i \neq j \), the edge \((i, j)\) is 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).

Hierarchical models are examples of discrete exponential families, see Barndorff-Nielsen (1978); Fienberg (2007); Rauh et al. (2011). Generalizing (3), let \( I \) and \( J \) be finite sets and let \( A \in \mathbb{R}^{J \times I} \) 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
\[
(4) \quad p_\theta(i) = \exp \left\{ \langle \theta, f_i \rangle - k(\theta) \right\} = \exp \left\{ (A^t \theta)_i - k(\theta) \right\}, \quad \theta \in \mathbb{R}^J,
\]
where, as above, \( k(\theta) = \log \sum_i \exp(\sum_j \theta_j a_{ji}) \). It is convenient to write \( \tilde{A} \) for the \((1 + |J|) \times I \) matrix with columns equal to \( (1, f_i) \), \( i \in I \), and to set \( \theta_0 := -k(\theta) \) and \( \tilde{\theta} = (\theta_0, \theta) \) (as a column vector). Then (4) rewrites to
\[
(5) \quad p_\theta(i) = \exp \left( \tilde{A}^t \tilde{\theta} \right), \quad \theta \in \mathbb{R}^J.
\]
Both \( A \) and \( \tilde{A} \) are called design matrices of the model. The convex hull of the columns \( f_i, i \in I \), is called the convex support polytope, denoted by \( P_A \). In the case of a graphical or hierarchical model, \( P_\Delta := P_{A_\Delta} \) is called a marginal polytope.

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

Later, the following reparametrization will be useful: select an element of \( I \), which we will denote by 0. Let \( A_0 \) be the matrix with columns \( f_i - f_0 \), \( i \in I \setminus \{0\} \). It is not difficult to see that \( A \) and \( A_0 \) define the same exponential family (since \( \tilde{A} \) and \( \tilde{A}_0 \) have the same row span). Let \( h' = \text{rank}(A_0) = \text{rank}(A_0) - 1 \), and select a set \( L \) of \( h' \) linearly independent vectors among the columns of \( A_0 \). For \( i \in L \), let \( \mu_i = \mu_i(\theta) := \langle \theta, f_i - f_0 \rangle \), and let \( \mu_L = (\mu_i, i \in L) \). Then the \( \mu_L \) are identifiable parameters on \( \mathcal{E}_A \); in fact, their number is equal to \( h' \), and they are independent by construction.

It is possible to extend the definition of \( \mu_i(\theta) \) to all \( i \in I \). Note that only the parameters \( \mu_i \) with \( i \in L \) are free parameters, while the parameters
\( \mu_i \) with \( i \in I \setminus L \) are linear functions of \( \mu_L \). The \( \mu_i \) can be interpreted as log-likelihood ratios:

\[
\mu_i(\theta) = \log \frac{p_\theta(i)}{p_\theta(0)}, \quad \mu_0(\theta) = 0.
\]

Let \( n = (n(i), i \in I) \) be an \( I \)-dimensional column vector of cell counts summarizing the outcome of a statistical experiment. Then

\[
(6) \quad \tilde{A}n = \begin{pmatrix} N \\ t \end{pmatrix} \quad \text{and} \quad An = t,
\]

where \( N = \sum_{i \in I} n(i) \) is the total cell counts and \( t \) is the column vector of sufficient statistic. The likelihood \( \prod_{i \in I} p_\theta(i)^{n(i)} \) can be written under the form of a natural exponential family. Indeed,

\[
\prod_{i \in I} p_\theta(i)^{n(i)} = \exp \left( \langle \tilde{A}n, \tilde{\theta} \rangle \right) = \exp \left\{ \sum_{j \in J} \theta_j t_j - Nk(\theta) \right\}.
\]

The log-likelihood function for the loglinear parameters \( \theta \) of \( E_A \) is therefore

\[
(7) \quad l(\theta) = \sum_{j \in J} \theta_j t_j - Nk(\theta).
\]

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 \( \mu = (\mu_i, i \in I) \):

\[
(8) \quad l(\mu) = \sum_{i \in I} n(i) \log p(i) = \sum_{i \in I} n(i) \mu_i - N \log(\sum_{i \in I} \exp \mu_i).
\]

As stated before, only a subset \( \mu_L \) of the parameters \( \mu \) are independent, and the remaining \( \mu_i, i \notin L \), can be expressed as linear functions of \( \mu_L \).

2.2. The convex support and its facial sets. We next recall some facts about facial sets. We refer to Ziegler (1998) for a general introduction to polytopes and their face lattices.

The convex support polytope \( P_A \) is defined as the convex hull of a finite number of points \( f_i, i \in I \). It is of interest to know which subsets of \( \{f_i\}_{i \in I} \) lie on a given face \( F \). Thus, we describe a face \( F \) by identifying the corresponding facial set \( F = \{i \in I : f_i \in F\} \). For any subset \( S \subseteq I \), denote by \( F_A(S) \) the smallest facial set that contains \( S \). The intersection of facial sets is again facial, and so \( F_A(S) \) is well-defined. When \( P_A = P_\Delta \) is a marginal polytope, we abbreviate \( F_A(\Delta)(S) \) by \( F_\Delta(S) \).
As mentioned in the introduction, to derive the inner approximation $F_1$ to $F_t$ and its outer approximation $F_2$, we need to consider sub-models of a given model. When one exponential family $E_{A'}$ is a subset of another family $E_A$, then the convex support polytope $P_{A'}$ is a linear projection of $P_A$, and the columns $f'_i$ of $A'$ are indexed by the same set $I$ as the columns $f_i$ of $A$. Since inverses of linear projections preserve faces, it follows from basic results about polytopes that $F_A(S) \subseteq F_{A'}(S)$; see Chapter 1 in Ziegler (1998). For hierarchical models, these facts are summarized in the following result:

**Lemma 2.1.** Let $\Delta$ and $\Delta'$ be simplicial complexes on the same vertex set with $\Delta' \subseteq \Delta$. Then $P_{\Delta'}$ is a coordinate projection of $P_{\Delta}$. The inverse image of any face of $P'$ is a face of $P$. Moreover, for any $S \subseteq I$, we have $F_{\Delta}(S) \subseteq F_{\Delta'}(S)$.

**Remark 2.2.** It is convenient to embed $P_A$ in a vector space with one additional dimension using a map $R^h \to R^{h+1}, t \mapsto (1, t)$. This has the advantage that all defining inequalities are brought into a homogeneous form with vanishing constant: 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 $F$ is given, its facial set $F$ can be obtained by checking whether $f_i \in F$ 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 $F$. However, it is straightforward to compute the linear equations defining $F$: the set of such equations $0 = \langle g, x \rangle - c = \langle \tilde{g}, \tilde{x} \rangle$ corresponds to the set of vectors $\tilde{g} \in \ker \tilde{A}_F$, where $\tilde{A}_F$ is the matrix obtained from $\tilde{A}$ by dropping the columns not in $F$.

**2.3. The closure of an exponential family and existence of the MLE.** For a family $E_A$ and cell counts $n = (n(i) : i \in I)$ given as above, a parameter value $\theta^*$ is an MLE if it is a global maximum of $l(\theta)$. An MLE need not exist, since the domain of the parameters $\theta$ is unbounded. The likelihood can also be written as a function of cell probabilities. For any probability distribution $p$ on $I$ let

$$\tilde{l}(p) = \log \left\{ \prod_{i \in I} p(i)^{n(i)} \right\}.$$  

Then $l(\theta) = \tilde{l}(p_{\theta^*})$, and $\theta^*$ is an MLE if and only if $p_{\theta^*}$ maximizes $\tilde{l}$ subject to the constraint that $p$ belongs to $E_A$, i.e. is of the form (4). When $\tilde{l}$ has no maximum on $E_A$, we can pass to the topological closure $\mathring{E}_A$. It can be characterized in terms of the convex support polytope $P_A$ and its facial sets as follows:
Theorem 2.3 (Barndorff-Nielsen (1978)). The topological closure of $E_A$ is $\overline{E_A} = \bigcup_F E_{F,A}$, where $F$ runs over all facial sets of $P_A$ and where $E_{F,A}$ consists of all probability distributions of the form $p_{F,\theta}$, with

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

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

Proof. See Barndorff-Nielsen (1978). For self-containedness we provide a proof in our notation in Appendix G.1. \hfill \square

Thus, $\overline{E_A}$ is a finite union of sets $E_{F,A}$ that are exponential families themselves with a very similar parametrization, using the same number of parameters. The design matrix of $E_{F,A}$ is the submatrix $A_F$ of $A$ consisting of the columns indexed by $F$. However, for any proper facial set $F \neq I$, the parametrization $\theta \mapsto p_{F,\theta}$ is never identifiable since all columns of $A_F$ lie on a supporting hyperplane defining $F$ and thus $A_F$ never has full rank.

Although the parameters $\theta$ on $E_A$ and the parameters $\theta$ on $E_{F,A}$ 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 $\theta$, then, in general, $\lim_{s \to \infty} \theta_j^{(s)} \neq \theta_j$ for all $j \in J$.

Theorem 2.4 (Barndorff-Nielsen (1978)). For any vector of observed counts $n$, there is a unique maximum $p^*$ of $\tilde{l} = \ln E_A$. This maximum $p^*$ satisfies: (1) $Ap^* = \frac{t}{N}$, where $t = An$, (2) $\text{supp}(p^*) = F_t$, (3) $p^* \in E_{F_t,A}$.

Proof. See Barndorff-Nielsen (1978). For self-containedness we provide a proof in our notation in Appendix G.2. \hfill \square

The maximum $p^*$ in Theorem 2.4 is called the extended maximum likelihood estimate (EMLE). By Theorem 2.4, when $F_t$ is known, the EMLE $p^*$ can be computed by computing the MLE on $E_{F_t,A}$. If the MLE $\theta^*$ exists, then $p^* = p_{\theta^*}$.

2.4. Decomposable models. Computing $F_t$ or finding an approximation is easier when the simplicial complex $\Delta$ of the model is decomposable. We need the following definitions.

Let $V' \subset V$. The restriction or induced subcomplex to $V'$ is $\Delta|_{V'} = \{S \in \Delta \mid S \subseteq V'\}$. The subcomplex $\Delta|_{V'}$ is complete, if $\Delta|_{V'}$ contains $V'$ (and thus all subsets of $V'$). In this case, we also say that $V'$ is complete in $\Delta$. 
A subset $S \subset V$ is a separator of $\Delta$ 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.

A hierarchical model is decomposable if its generating set is 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 $\Delta_i$ 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, $\Delta$ arises by iteratively gluing simplices along complete sub-simplices.

Lemma 2.5 below states that, if $\Delta$ is reducible, then any facial set for $\Delta$ is the intersection of the preimage of facial sets for its components. It is a simple reformulation of Lemma 8 in (Eriksson et al., 2006).

**Lemma 2.5.** Let $\Delta$ be reducible into two components $\Delta|_{V_1}$ and $\Delta|_{V_2}$.

1. If $F \subseteq I$ is facial with respect to $\Delta$, 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 $\Delta$.

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

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

Lemma 2.5 generalizes to more than one separator and thus to more than two components. It becomes particularly simple when these components are complete: in that case, $F_{\Delta|_{V_1}}(T_1) = T_1$. Taking the preimage we obtain

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

Thus, for a decomposable complex $\Delta = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_r$, we have

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

for any $T \subseteq I$, where $\pi_i = \pi_{V(\Delta_i)}$.

**3. Approximations of facial sets.** We consider a hierarchical model with simplicial complex $\Delta$ and marginal polytope $P_\Delta$. In this section, we develop the details of our methodology to obtain inner and outer approximations to the facial set $F_t$ of the data vector $t$. 
3.1. Inner approximations. To obtain an inner approximation, our strategy is to find a separator $S$ of $\Delta$ and to complete it. To be precise, we augment $\Delta$ by adding all subsets of $S$. Thus, we obtain a simplicial complex $\Delta_S = \Delta \cup \{M : M \subseteq S\}$ in which $S$ is a complete separator. We can apply Lemma 2.5 to find the facial set $F_{\Delta_S}(I_+)$, and this will be our inner approximation of $F_t$, because $F_{\Delta_S}(I_+) \subseteq F_{\Delta}(I_+) = F_t$ according to Lemma 2.1.

An even simpler approximation is obtained by not only completing the separator itself, but also the two parts $V_1, V_2$ separated by $S$: the simplicial complex $\Delta_{V_1,V_2} := \{M : M \subseteq V_1\} \cup \{M : M \subseteq V_2\}$ is decomposable and contains $\Delta$. Its facial sets can be computed from (10).

In general, the approximation obtained from a single separator (or, in general, a single super-complex) is not good; that is, $F_{\Delta_S}(I_+) = F_{\Delta_{V_1,V_2}}(I_+)$ tends to be much larger than $F_{\Delta_S}(I_+)$ or $F_{\Delta_{V_1,V_2}}(I_+)$. Thus we need to combine information from several separators. For example, given two separators $S, S' \subseteq V$, we find a chain of approximations

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

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 $\Delta_S$ or $\Delta_{S'}$. The last inclusion is a consequence of Lemma 2.1 since both $\Delta_S$ and $\Delta_{S'}$ contain $\Delta$. This chain of approximations has to stabilize; that is, after a certain number of iterations, the approximations will not improve any more. The limit $F_{S,S'}(I_+) := \bigcup_i G_i = \bigcup_i G'_i$ can be characterized as the smallest subset of $I$ that contains $I_+$ and is facial both with respect to $\Delta_S$ and $\Delta_{S'}$. The same iteration can be done replacing $\Delta_S$ by $\Delta_{V_1,V_2}$ and $\Delta_{S'}$ by $\Delta_{V'_1,V'_2}$. Applying in turn $F_{\Delta_{V_1,V_2}}$ and $F_{\Delta_{V'_1,V'_2}}$ gives another approximation $\tilde{F}_{S,S'}(I_+)$, namely the smallest subset of $I$ that contains $I_+$ and is facial both with respect to $\Delta_{V_1,V_2}$ and $\Delta_{V'_1,V'_2}$. This latter approximation will be used in Section 5.1. Clearly, $\tilde{F}_{S,S'}(I_+)$ is a worse approximation than $F_{S,S'}(I_+)$, since $\tilde{F}_{S,S'}(I_+) \subseteq F_{S,S'}(I_+) \subseteq F_t$, but it is easier to compute.

We use the following strategies:

1. if possible, use all separators of a graph.
We illustrate this strategy in Section 5.2, using a graphical model associated with the NLTCS data set.

There are two problems with this strategy: First, if \( S \) is such that either \( V_1 \) or \( V_2 \) is large, then it becomes difficult to compute \( F_{\Delta|V_1} \) and \( F_{\Delta|V_2} \). Such “bad” separators always exist: namely, each node \( i \in V \) is separated by its neighbours from all other nodes. In this case, \( V_1 \) consists of \( i \) and its neighbours, and \( V_2 \) consists of \( V \setminus \{i\} \). For such a “bad” separator we can only compute \( F_{\Delta|V_1,V_2} \), but not \( F_{\Delta|S} \). Second, the number of separators may be large. Thus, when computing the inner approximation, it may take a long time until the iteration over all separators converges. A faster alternative strategy is the following:

2. use all separators such that both \( V_1 \setminus S \) and \( V_2 \setminus S \) are not too small (for example, \( \min\{|V_1 \setminus S|,|V_2 \setminus S|\} \geq 3 \)).

In the case of the grids studied in Sections 5.1 and 6.2, which have a lot of regularity, we use an adapted strategy:

3. in a grid, use the horizontal, vertical and diagonal separators.

In the case of grids, the vertical separators form a family of pairwise disjoint separators. In Section 6 we show how to make use of such a family to study faces of hierarchical models, even if the facial sets are so large that they become computationally intractable.

3.2. Outer approximations. By Lemma 2.1, the facial set \( F_{\Delta'}(S) \) for a simplicial sub-complex \( \Delta' \subseteq \Delta \) provides an outer approximation of \( F_{\Delta}(S) \). Removing sets from \( \Delta \) decreases the dimension of the marginal polytope, so it is often easier to compute \( F_{\Delta'}(S) \) than to compute \( F_{\Delta}(S) \). Our main strategy is to look at induced sub-complexes.

When comparing \( \Delta \) with an induced sub-complex \( \Delta|_{V'} \) for some \( V' \subset V \), we have to be precise about whether we consider \( \Delta|_{V'} \) as a complex on \( V \) or on \( V' \). When we consider it on \( V \), then its design matrix \( A \) has columns \( f_i \) indexed by \( i \in I \). When we consider it on \( V' \), its design matrix \( A' \) has columns \( f'_i \) indexed by \( I_{V'} \). Because we have the same set of interactions whether we are on \( V \) or \( V' \), we have for \( i \in I \) and \( i' \in I_{V'} \),

\[
f_i = f'_i \iff i \in \pi_{V'}^{-1}(i').
\]

Therefore the marginal polytopes of the two models are the same since they are the convex hull of the same set of vectors \( \{f_i, i \in I\} = \{f'_i, i' \in I_{V'}\} \). The relationship between the facial sets on \( V \) and \( V' \) is as follows:
Lemma 3.1. Let $V' \subseteq V$. For $K \subseteq I$, we have
\[ F_{\Delta|_{V'}}(K) = \pi_{V'}^{-1}(F_{\Delta|_{V'}}'(\pi_{V'}(K))). \]

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. For $K \subseteq I$, the two sets $A = \{a_i, i \in K\}$ and $B = \{b_{i'}, i' \in \pi_{V'}(K)\}$ are identical and therefore the smallest faces of the marginal polytopes for $\Delta_{V'}$ on $V$ or $V'$ containing $A$ and $B$ respectively are the same.

By definition of $F'_{\Delta|_{V'}}(\pi_{V'}(K))$, the smallest face containing $B$ is defined by $\{b_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. By definition of $F_{\Delta_{V'}}(K)$, the smallest face containing $A$ is $\{a_i, i \in F_{\Delta_{V'}}(K)\}$. Also by (11), we have that $\{a_i, i \in \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))\} = \{b_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. Therefore $F_{\Delta_{V'}}(K) = \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))$.

In general, $F_{\Delta|_{V'}}(I_+)$ is not a good approximation of $F_{\Delta}(I_+)$. This approximation can be improved by considering several subsets of $V$. To be precise, if $V_1, \ldots, V_r \subseteq V$, then $F_{\Delta}(I_+) \subseteq F_{\Delta|_{V_i}}(I_+)$ for $i = 1, \ldots, r$, and thus $F_{\Delta}(I_+) \subseteq \bigcap_{i=1}^r F_{\Delta|_{V_i}}(I_+) =: F_{V_1, \ldots, V_r \Delta}(I_+)$. In contrast to the case of the inner approximation, no repeated iteration is needed. Thus, the outer approximation is faster to compute.

The question is now how to choose the subsets $V_i$. Clearly, the subsets $V_i$ should cover $V$, and, more precisely, they should cover $\Delta$, in the sense that for any $D \in \Delta$ there should be one $V_i$ with $D \in \Delta|_{V_i}$. The larger the sets $V_i$, the better the approximation becomes, but the more difficult it is to compute $F_{V_1, \ldots, V_r \Delta}(I_+)$. One generic strategy is the following:

1. use all subsets of $V$ of fixed cardinality $k$ plus all facets $D \in \Delta$ with $|D| \geq k$.

This choice of subsets indeed covers $\Delta$. The parameter $k$ should be chosen as large as possible such that computing $F_{V_1, \ldots, V_r \Delta}(I_+)$ is still feasible. Note that computing $F_{\Delta|_D}(I_+)$ for $D \in \Delta$ is trivial, since $P_{\Delta|_D}$ is a simplex. Another natural strategy, due to Massam and Wang (2015), is the following:

2. for fixed $k$, use balls $B_k(v) = \{w : d(v, w) \leq k\}$ around the nodes $v \in V$, where $d(\cdot, \cdot)$ denotes the edge distance in the graph.

Our general philosophy is that the subsets $V_i$ should be large enough to preserve some of the structure of $\Delta$. For example, for the grid graphs, we
suggest to use $3 \times 3$ sub-grids. These graphs have two nice properties: First, they already have the appearance of a small grid. Second, for any vertex $v \in V$, there is a $3 \times 3$ sub-grid that contains $v$ and all neighbours of $v$. We will compare two different strategies:

3. for a grid, use all $3 \times 3$ sub-grids;
4. cover a grid by $3 \times 3$ sub-grids.

In Section 6.2 we compare these two methods, and we observe that, in the example of the $5 \times 10$ grid, it suffices to only look at a covering.

In general, it is not enough to look at induced sub-complexes, unless $\Delta$ has a complete separator (see Section 2.4). However, the approximation tends to be good and gives the correct facial set in many cases.

3.3. 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.

The sets $F_1, F_2$ and $F_t$ can also be compared by the ranks of the matrices $\hat{A}_{F_1}, \hat{A}_{F_2}$ and $\hat{A}_{F_t}$ obtained from $\hat{A}$ by keeping only the columns indexed by $F_1, F_2$ and $F_t$, respectively. Clearly, $\text{rank } \hat{A}_{F_1} \leq \text{rank } \hat{A}_{F_t} \leq \text{rank } \hat{A}_{F_2}$. Note that $\text{rank } \hat{A}_{F_2} - 1$ equals the dimension of the corresponding face $F_2$ of $P$, and $\text{rank } \hat{A}_{F_t} - 1$ equals the dimension of $F_t$. Although $F_1$ does not necessarily correspond to a face of $P$, we can bound the codimension of $F_t$ in $F_2$ by

$$\dim F_2 - \dim F_t \leq \text{rank } \hat{A}_{F_2} - \text{rank } \hat{A}_{F_1}.$$ 

In particular, if $\text{rank } \hat{A}_{F_2} = \text{rank } \hat{A}_{F_1}$, 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. Parameter Estimation when the MLE does not exist.

4.1. Computing the extended MLE. If the MLE $\theta^*$ exists, then it can be computed by finding the unique maximum of the log-likelihood function $l(\theta)$ given in (7). As mentioned before, $l(\theta)$ is concave (or even strictly concave, if the parameters $\theta$ are identifiable), and thus the maximum is, at least in principle, easy to find (in practice, for larger models, it may be difficult to evaluate the function $k(\theta)$, which involves a sum over $I$; but we will not discuss this problem here). In general, the maximum cannot be found symbolically, but there are efficient numerical algorithms to maximize concave functions. Any reasonable hill-climbing algorithm should be capable
of finding the MLE. An example of an algorithm commonly used is *iterative proportional fitting* (IPF), which can be thought of as an algorithm of Gauss-Seidel type (Csiszár and Shields, 2004).

When the MLE does not exist but the facial set $F = F_t$ of the data is known, then it is straightforward to compute the extended MLE $p^*$. In this case, we know that $p^*$ lies in $\mathcal{E}_{F,A}$. To find $p^*$, we need to optimize the log-likelihood $\tilde{l}$ over $\mathcal{E}_{F,A} = \{p_F, \theta : \theta \in \mathbb{R}^h\}$. Plugging the parametrization $p_F, \theta$ (see Theorem 2.3) into $\tilde{l}$ tells us that we need to optimize the restricted log-likelihood function

$$l_F(\theta) = \log \left( \prod_{i \in I_+} p_F, \theta(i)^{n(i)} \right) = \sum_{j \in J} \theta_j t_j - N k_F(\theta).$$

This problem is of a similar type as the problem to maximize $l$ in the case that the MLE exists, and the same algorithms as discussed above can be used. The problem here is slightly easier, since $F$ is smaller than $I$. However, as stated above, the parametrization $\theta \mapsto p_F, \theta$ is never identifiable. Of course, this problem is easy to solve by selecting a set of independent parameters among the $\theta_j$. However, depending on the choice of the independent subset, the values of the parameters change, and in particular, it is meaningless to compare the values of the parameters $\theta_j$ with parameter values of any other distribution in $\mathcal{E}_A$ or in the closure $\overline{\mathcal{E}_A}$.

Before explaining how to find better parameters on $\mathcal{E}_{F,A}$, let us discuss what happens if the facial set $F_t$ of the data is not known. As mentioned before, whether or not the MLE exists, the log-likelihood function $l(\theta)$ is always strictly concave (assuming that the parametrization is identifiable). When the MLE does not exist, then the maximum is not at a finite value $\theta^*$, but lies “at infinity.” Still, as observed by Geyer (2009, Section 3.15), any reasonable numerical “hill-climbing” algorithm that tries to maximize the likelihood will tend towards the right direction. Such a numeric algorithms generates a sequence of parameter values $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ with increasing log-likelihood values $l(\theta^{(1)}) \leq l(\theta^{(2)}) \leq \ldots$. Since $l(\theta)$ is concave, our optimization problem is numerically easy (at least in theory), and for any reasonable such algorithm, the limit $\lim_{s \to \infty} l(\theta^{(s)})$ will equal $\sup_{\theta \in \mathcal{E}_A} l(\theta) = \max_{p \in \mathcal{E}_A} \tilde{l}(p)$. The algorithm will stop when the difference $l(\theta^{(s+1)}) - l(\theta^{(s)})$ becomes negligibly small. The output, $\theta^{(s)}$, then gives a good approximation of the EMLE, in the sense that $p^*$ and $p_{\theta^{(s)}}$ are close to each other. For many applications, such as in machine learning, where it is more important to have good values of the parameters instead of trying to model the “true underlying distribution,” or when doing a likelihood test, where the value of the likelihood is more important than the parameter values, this may be good enough.
However, in this numerical optimization, some of the parameters $\theta_j$ will tend to $\pm\infty$, which may lead to numerical problems. For example, it may happen that one parameter goes to $+\infty$ and a second parameter to $-\infty$ in such a way that their sum remains finite (see Appendix B for a simple such example with two variables). This implies that a difference between two large numbers has to be computed, which is numerically unstable. Also, it is not clear, which parameters tend to infinity numerically. In fact, this may depend on the chosen algorithm; i.e. different algorithms may yield approximations of the EMLE that are qualitatively different in the sense that different parameters diverge.

To avoid such problems, we propose a change of coordinates that allows us to control which parameters diverge, at least in the case where we know the facial set $F_t$. If $F_t$ is unknown, but if we know approximations $F_1 \subseteq F_t \subseteq F_2$, we can use this knowledge to identify some parameters that definitely remain finite, while some parameters definitely diverge. We cannot control the behaviour of the remaining parameters, but, as will be illustrated in Section 5.2, the MLE obtained with the model on $F_2$ lies closer to the EMLE than the MLE on the original model. The more information we have about the facial set $F_t$, the better we can control the above mentioned pathologies.

4.2. An identifiable parametrization. We have seen that when we use the parametrization $\theta \mapsto p_{F_t,\theta}$ of $E_{A,F_t}$ in the case where $F_t \neq I$, we have to expect the following (interrelated) issues:

1. The parametrization is not identifiable, i.e. there are parameters $\theta, \theta'$ with $p_{F_t,\theta} = p_{F_t,\theta'}$.
2. While the parametrization $\theta \mapsto p_{F_t,\theta}$ of $E_{F_t,A}$ looks similar to the parametrization $\theta \mapsto p_{\theta}$ of $E_{A}$, the values of the parameters in both parametrizations are not related to each other.
3. When $p_{\theta(s)} \rightarrow p_{F_t,\theta}$ as $s \rightarrow \infty$ for some parameter values $\theta(s), \theta$, then some of the parameter values $\theta(s)$ diverge to $\pm\infty$. When computing probabilities, there may be linear combinations of these diverging parameters that remain finite.

Next we show that if $F_t$ is known, then, with a convenient choice of $L$, the parameters $\mu_L$ (introduced in Section 2.1) solve 1 and 2 and improve 3. Afterwards, we discuss what can be done if $F_t$ is not known. We briefly discuss the general solution towards 3 in Appendix C. In any case, the choice of the parameters will depend on the facial set $F_t$: it is not possible to define a single parametrization that works for all facial sets simultaneously.

Suppose that $F_t$ is known. We choose a zero element in $I_+$ and consider
the parameters $\mu_i$ as in Section 2. Recall that

$$\mu_i(\theta) = \langle \theta, f_i - f_0 \rangle = \log p(i) / p(0), \ i \in I.$$  

As mentioned in Section 2, the parameters $\mu_i$ are not independent, and we need to choose an independent subset $L$. We will do this in two steps.

1. Choose a maximal subset $L_t$ of $F_t$ such that the parameters $\mu_i, \ i \in L_t$ are independent.

2. Then extend $L_t$ to a maximal subset $L \subseteq I$ such that the parameters $\mu_i, i \in L$ are independent by adding elements $i \in I \setminus F_t$.

It follows from Theorem 2.4 that the following holds.

1. The subset $\mu_i, \ i \in L_t$, of the parameters $\mu_L$ gives an identifiable parametrization of $E_{F_t, A}$.

2. Let $\mu^*_i, \ i \in L_t$, be the parameter values that maximize $l_{F_t}$ (and thus give the EMLE). When the likelihood $l(\mu)$ in (8) is maximized numerically on $I$, then in successive iterations of the maximization, the estimates $\mu^{(s)}_i$ are such that

$$\mu^{(s)}_i \to \begin{cases} \mu^*_i, & i \in L_t, \\ -\infty, & \text{otherwise}. \end{cases}$$

In particular, no parameter tends to $+\infty$.

The last property ensures a consistency of the parameters $\mu_i$ on $E_A$ and on $E_{F_t, A}$. This is important in those cases where the parameters have an interpretation and where it is of interest to know the value of some parameters, if it is well-defined. For example, in hierarchical models, the parameters correspond to “interactions” of the random variables, and it may be of interest to know, which of these interactions are important. Thus, it is of interest to know the size of the corresponding parameter. Usually, it is not the parameter $\mu_i$, but the original parameters $\theta_i$ that have an interpretation. But when we understand the parameters $\mu_i$, we can also tell which of the parameters $\theta_i$ or which combinations of the parameters $\theta_i$ have finite well-defined values and can be computed, and which parameters diverge:

**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

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

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 a linear-combination of the $\mu^{(s)}_i$ with $i \in L_t$. 


Proof. The first statement follows from
\[
\mu_i^{(s)} = \log p_{\theta^{(s)}}(i)/p_{\theta^{(s)}}(0) \to \log p^*(i)/p^*(0).
\]

For the second statement, note that any linear combination of the \( \theta \) is also a linear combination of the \( \mu \), since the linear map \( \theta \mapsto \mu(\theta) \) is invertible.

We now show that if a linear combination \( \sum_i a_i \mu_i \) involves some \( \mu_j \) with \( j \notin L \), then there exist sequences \( \mu^{(s)}_i \), \( \mu'^{(s)}_i \) of parameters with
\[
\lim_{s \to \infty} p_{\mu^{(s)}_i} = \lim_{s \to \infty} p_{\mu^{(s)}_i} \quad \text{and} \quad \lim_{s \to \infty} \sum_i a_i \mu^{(s)}_i \neq \lim_{s \to \infty} \sum_i a_i \mu'^{(s)}_i.
\]

So suppose that \( \mu^{(s)} \) is a sequence of parameters such that \( \lim_{s \to \infty} p_{\mu^{(s)}_i} \) exists and such that \( \lim_{s \to \infty} \sum_i a_i \mu^{(s)}_i \) is finite. Define
\[
\mu'^{(s)}_i = \begin{cases} 
\mu^{(s)}_i + 1, & \text{if } i = j, \\
\mu^{(s)}_i, & \text{otherwise}.
\end{cases}
\]

An easy computation shows that
\[
\lim_{s \to \infty} p_{\mu'^{(s)}_i} = \lim_{s \to \infty} p_{\mu^{(s)}_i} \quad \text{and} \quad \lim_{s \to \infty} \sum_i a_i \mu'^{(s)}_i = \lim_{s \to \infty} \sum_i a_i \mu^{(s)}_i + a_j. \quad \blacksquare
\]

Suppose now that we do not know \( F_t \), but that instead we have approximations \( F_1, F_2 \) that satisfy
\[
I_+ \subseteq F_1 \subseteq F_t \subseteq F_2 \subseteq I.
\]

In this case, we proceed as follows to obtain an independent subset \( L \) among the parameters \( \mu_i \):

1. Choose a maximal subset \( L_1 \) of \( F_1 \) such that the parameters \( \mu_i, i \in L_1 \) are independent.
2. Then extend \( L_1 \) to a maximal subset \( L_2 \subseteq F_2 \) such that the parameters \( \mu_i, i \in L_2 \) are independent by adding elements \( i \in F_2 \setminus F_1 \).
3. Finally, extend \( L_2 \) to a maximal subset \( L \subseteq I \) such that the parameters \( \mu_i, i \in L \) are independent by adding elements \( i \in I \setminus F_2 \).

These parameters have the following properties that follow directly from Lemma 4.1:

**Lemma 4.2.** Suppose that \( \theta^{(s)}, s \in \mathbb{N} \), are parameter values such that \( p_{\theta^{(s)}} \to p^* \) as \( s \to \infty \), and let \( \mu^{(s)}_i = \langle \theta^{(s)}, f_i \rangle \).
1. For any $i \in L_1$, the linear combination 
\[ \mu_i^{(s)} = \langle \theta, f_i \rangle \]
has a well-defined finite limit as $s \to \infty$. Thus, any linear combination of the $\mu_i^{(s)}$ with $i \in L_1$ has a well-defined limit as $s \to \infty$.

2. Any linear combination $\sum_i a_i \mu_i^{(s)}$ that has a well-defined limit as $s \to \infty$ is in fact a linear combination of the $\mu_i^{(s)}$ with $i \in L_2$. Thus, a linear combination that involves at least one $\mu_j^{(s)}$ with $j \in L \setminus L_2$ does not have a well-defined limit.

5. Simulation study and applications to real data. In this section, we illustrate our methodology. In 5.1, we simulate data for the graphical model of the $4 \times 4$ grid and show how to exploit the various types of separators in order to obtain good inner and outer approximations. We find that our method gives very accurate result in this model of modest size. In 5.2, we work with the NLTCS data set, a real-world data set. We compare different inner approximations $F_1$ and find that most of the time, $F_1$ and $F_2$ are equal, and thus they are both equal to $F_t$. We also compute the EMLE and compare these exact estimates to those obtained when maximizing the likelihood functions $l$ and $l_{F_2}$. We find the results given by $l_{F_2}$ better than those given by $l$, and extremely close to the finite components of the EMLE.

5.1. $4 \times 4$ grid graph. We generated random samples of varying sizes for the graphical model of the $4 \times 4$ grid graph with binary variables (Fig. 1a). For each sample, we compute inner and outer approximations $F_1$ and $F_2$, and we compare them to the true facial set $F_t$, which we can obtain using linear programming. To obtain an inner approximation, we use two strategies. Either, we iterate over all possible separators, of which there are 106 (Strategy (1) in Section 3.1), or we iterate over the 3 horizontal, 3 vertical and 8 diagonal separators only (Strategy (3) in Section 3.1). We obtain the same result with either strategy. Clearly, Strategy (3) is much faster. To compute the outer approximation, we cover the $4 \times 4$ grid by four $3 \times 3$ grids (Strategy (3) in Section 3.2).

We generate samples from the hierarchical model $P_\theta$, where the vector of parameters $\theta$ is drawn from a multivariate standard normal distribution (for each sample, new parameters were drawn). The results are given in Table 1. For each sample size, 1000 samples were obtained. Observe that the squared length of the parameter vector $\theta$ is $\chi^2$-distributed with 40 degrees of freedom (since the number of parameters is 40). Thus, the expected length of $\theta$ is 40, which is large enough to move the distribution $p_\theta$ close to the boundary of the
model. Indeed, we observed that when the MLE does not exist, the length of the numerical estimate of the MLE vector is of the order of magnitude of 40 (see also the next example in Section 5.2). In all samples that we generated, $F_1 = F_2$, and $F_1 = F_2$ in the vast majority of cases. Thus, for this graph of relatively modest size, our approximations are very good. We present additional simulation results in Appendix D.

5.2. NLTCS data set. To illustrate how approximate knowledge of the facial set allows us to say which parameters can be estimated (as explained in Section 4), we study the NLTCS data set, which consists of 21,574 observations on 16 binary variables, called ADL1, . . . , ADL6, IADL1, . . . , IADL10. The reader is referred to Dobra and Lenkoski (2011) for a detailed description of the data set. To associate a hierarchical model to this data, we rely on the results of Dobra and Lenkoski (2011) who use a Bayesian approach to estimate the posterior inclusion probabilities of edges. We construct a graph by saying that $(x, y)$ is an edge if and only if the posterior inclusion probability of $(x, y)$ is at least 0.40: we obtain Figure 1b. Then we take the corresponding clique complex of this graph so that our hierarchical model is a graphical model. There are 314 parameters in this model, including up to 6-way interactions. In total, the graph has 40 separators.

In order to compare the estimates obtained with or without worrying about the existence of the MLE and with or without an approximation to $F_1$, we maximize the loglikelihood given in terms of $\mu$, rather than $\theta$, as in (8). First we ignore the fact that the MLE might not exist and numerically optimize the likelihood directly: we call this estimate $\hat{\mu}^{\text{MLE}}$. Second, we find $F_1$ and compute the EMLE with parameters denoted $\hat{\mu}^{\text{EMLE}}$. Third, we obtain an inner and outer approximation to $F_1$ and consider the resulting information on likelihood maximization. We call the resulting estimate $\hat{\mu}^{F_1\setminus F_1'}$. All estimates are computed using the Matlab function minFunc (Schmidt, 2005).

To compute $\hat{\mu}^{\text{EMLE}}$, we first compute the inner approximation $F_1$ that
makes use of all the separators in the graph (Strategy 1 in Section 3.1). We also compute an outer approximation $F_2$ from all $\binom{16}{5} = 4368$ size five local models and the cliques of size six (Strategy 1 in Section 3.2). We obtain $F_1 = F_2$ and thus deduce that $F_t = F_1 = F_2$. We find $|F_t| = 49536$, and so $|F^c_t| = 2^{16} - 49536 = 16000$. Therefore, 16000 cell probabilities are zero in the EMLE, a precise estimate of those cells that we could not obtain from the MLE. We obtain the EMLE by maximizing the loglikelihood function $l_{F_t}$ as in (12). Since $\text{rank}(\tilde{A}_{F_t}) = 303$, the dimension of $F_t$ is 302, and there are only 302 parameters in $l_{F_t}$. This information is most important when testing the present model against another model $\mathcal{M}_2$ of smaller dimension.

As pointed out by Geyer (2009) and Fienberg and Rinaldo (2012), the test statistic, chi-square or loglikelihood, has to be compared to the chi-square distribution with $302 - d_2$ degrees of freedom, not $314 - d_2$. Of course, for $\mathcal{M}_2$ also, $d_2$ is the dimension of the smallest face of the corresponding polytope containing the data.

To show how to use the inner and outer approximations when $F_t$ is not known, we construct coarser inner and outer approximations to $F_t$, respectively denoted $F'_1$ and $F'_2$, and use them to compute another approximation $\hat{\mu}^{F'_2 \setminus F'_1}$ to the EMLE. To compute $F'_1$, we just use 10 random separators. We
find \(|F'_1| = 36,954\) and \(\dim F'_1 = \text{rank } \tilde{A}_{F'_1} - 1 = 300\). To compute the outer approximation \(F'_2\), we consider the 4368 local size-five induced models and select among them the 1000 with the facial sets of smallest cardinality, which we glue together. We find \(|F'_2| = 50,688\) and \(\dim F'_2 = \text{rank } \tilde{A}_{F'_2} - 1 = 310\). Thus, we know that at least \(|I \setminus F'_2| = 2^{16} - 50,688 = 14,848\) cell probabilities vanish in the EMLE. Since we pretend not to know \(F_t\), we replace \(l_{F_t}\) by

\[
 l_{F'_2}(\mu) = \sum_{i \in I_+} \mu_i n(i) - N \sum_{i \in F'_2} \exp(\mu_i).
\]

We know that \(\mu_i\) is estimable for \(i \in F'_1\), that \(\mu_i\) goes to negative infinity for \(i \in F'_2^c\), and we cannot say anything for \(\mu_i\) with \(i \in F'_2 \setminus F'_1\).

As explained in Section 4.2, the components of \(\mu\) are not functionally independent. We choose \(L_1 \subseteq F'_1\), \(L_2 \subseteq F'_2\) and \(L \subseteq I\) as in Section 4.2 (we note that the zero cell belongs to \(I_+\)). Then any \(\mu_i\), \(i \in F'_t\), can be written as a linear combination of \(\mu_{L_2} = (\mu_i, i \in L_2)\), and we can write \(\mu_i = \langle b_i, \mu_{L_2} \rangle\) for an appropriate vector \(b_i\). Thus, \(l_{F'_2}(\mu)\) only depends on \(\mu_{L_2} = (\mu_i, i \in L_2)\), and (13) can be rewritten as

\[
 l_{F'_2}(\mu_{L_2}) = \sum_{i \in I_+} \langle b_i, \mu_{L_2} \rangle n(i) - N \sum_{i \in F'_2} \exp(b_i, \mu_{L_2}).
\]

Of course, the maximum of \(l_{F'_2}\) does not exist but, as for the maximization of \(l\), the computer still gives us a numerical approximation, \(\hat{\mu}_{L_2}\), and thus also a numerical estimate \(\hat{\mu}_i = \langle b_i, \hat{\mu}_{L_2} \rangle\), \(i \in F'_2\). In total, there are \(|L_2| = \text{rank } \tilde{A}_{F'_2} - 1 = 310\) independent parameters in the loglikelihood function (14). Among them, there are \(|L_1| = \text{rank } \tilde{A}_{F'_1} - 1 = 300\) estimable parameters \(\mu_i, i \in L_1\). We cannot say anything about the 10 parameters indexed by \(L_2 \setminus L_1\). If we know \(F_t\), we can identify two more estimable parameters.

Table 2 gives the three estimates \(\hat{\mu}_i^{\text{MLE}}\), \(\hat{\mu}_i^{\text{EMLE}}\) and \(\hat{\mu}_i^{F'_2 \setminus F'_1}\). The naive estimator \(\log \frac{n_i}{n_0}\) is also listed. Estimates are listed for 19 arbitrarily chosen parameters among the 310 possible ones. The first column of the table indicates whether the index \(i\) belongs to \(F'_1\), \(F_t\) or \(F'_2\). The second column lists the particular parameters considered. By Theorem 2.4, the only parameters \(\mu_i\) with a finite estimate are those for \(i \in F_t\). This is illustrated in the \(\hat{\mu}^{\text{EMLE}}\) column of Table 2, with finite values for \(\hat{\mu}_i^{\text{EMLE}}\), \(i \in F_t\) (green and pink rows), and infinite values for \(\hat{\mu}_i^{\text{EMLE}}\), \(i \in I \setminus F_t\) (yellow and blue rows).

When we choose the coarser approximations, \(F'_t\) and \(F'_2\), to \(F_t\), we compute the estimate \(\hat{\mu}^{F'_2 \setminus F'_1}\) using (14). The components \(\hat{\mu}_i^{F'_2 \setminus F'_1}\) indexed by \(i \in F_t\) are excellent. They are finite and close to the corresponding
Table 2

Parameter estimates from 3 methods compared with the relative frequency in the NLTCS data. Here, each $i = (s_1, \ldots, s_{16}) \in I = \{0, 1\}^{16}$ is represented by the natural number $\sum_{j=1}^{16} s_j 2^{j-1} \in \{0, \ldots, 2^{16} - 1\}$.

| Parameter | naive estimate | maximum likelihood estimates |
|-----------|----------------|-----------------------------|
|           | $\log n_i/n_0$ | $\hat{\mu}_{i}^{\text{MLE}}$ | $\hat{\mu}_{i}^{\text{EMLE}}$ | $\hat{\mu}_{i}^{F_2 \setminus F_1}$ |
| $i \in F_1$ | $\hat{\mu}_{512}$ | $-1.2472$ | $-1.2482$ | $-1.2482$ |
|           | $\hat{\mu}_{6536}$ | $-1.7644$ | $-1.7976$ | $-1.7975$ |
|           | $\hat{\mu}_{116}$  | $-2.3958$ | $-2.3844$ | $-2.3846$ |
|           | $\hat{\mu}_{528}$  | $-2.5429$ | $-2.6504$ | $-2.6504$ |
|           | $\hat{\mu}_{2048}$ | $-2.8813$ | $-2.7246$ | $-2.7243$ |
| $i \in F_2 \setminus F_1$ | $\hat{\mu}_{32960}$ | $-\infty$ | $-13.8205$ | $-13.8207$ |
|           | $\hat{\mu}_{34881}$ | $-\infty$ | $-14.3693$ | $-14.3693$ |
| $i \in F_2$ | $\hat{\mu}_{36864}$ | $-\infty$ | $-30.8729$ | $-34.9805$ |
|           | $\hat{\mu}_{36880}$ | $-\infty$ | $-39.6536$ | $-45.2229$ |
|           | $\hat{\mu}_{388}$   | $-\infty$ | $-28.9090$ | $-29.4525$ |
|           | $\hat{\mu}_{32769}$ | $-\infty$ | $-32.3799$ | $-36.9537$ |
|           | $\hat{\mu}_{385}$   | $-\infty$ | $-37.1365$ | $-35.9399$ |
|           | $\hat{\mu}_{449}$   | $-\infty$ | $-38.9673$ | $-44.9405$ |
|           | $\hat{\mu}_{32785}$ | $-\infty$ | $-40.1221$ | $-45.8318$ |
|           | $\hat{\mu}_{389}$   | $-\infty$ | $-43.7297$ | $-40.0158$ |
| $i \in I \setminus F_2$ | $\hat{\mu}_{256}$  | $-\infty$ | $-35.5482$ | $-\infty$ |
|           | $\hat{\mu}_{320}$   | $-\infty$ | $-42.5454$ | $-\infty$ |
|           | $\hat{\mu}_{257}$   | $-\infty$ | $-52.9224$ | $-\infty$ |
|           | $\hat{\mu}_{321}$   | $-\infty$ | $-60.2208$ | $-\infty$ |

of $\hat{\mu}_{i}^{\text{EMLE}}$. This can be seen by verifying numerically that the square length of the projection on $F_1$ of the difference between $\hat{\mu}_{i}^{\text{MLE}}$ and $\hat{\mu}_{i}^{\text{EMLE}}$ is greater than that between $\hat{\mu}_{i}^{F_2 \setminus F_1}$ and $\hat{\mu}_{i}^{\text{EMLE}}$. Indeed, we have

$$
\|\hat{\mu}_{i}^{F_2 \setminus F_1} - \hat{\mu}_{i}^{\text{EMLE}}\|^2 \approx 6.49 < \|\hat{\mu}_{i}^{\text{MLE}} - \hat{\mu}_{i}^{\text{EMLE}}\|^2 \approx 8.52.
$$

The components $\hat{\mu}_{i}^{F_2 \setminus F_1}$ indexed by $i \in F_2 \setminus F_1$ are finite while the corresponding components of $\hat{\mu}_{i}^{\text{EMLE}}$ are infinite but they are better than those of $\hat{\mu}_{i}^{\text{MLE}}$: numerically, we have

$$
\sum_{i \in F_2 \setminus F_1} (\hat{\mu}_{i}^{F_2 \setminus F_1})^2 \approx 5184 > \sum_{i \in F_2 \setminus F_1} (\hat{\mu}_{i}^{\text{MLE}})^2 \approx 4752.
$$

The estimates $\hat{\mu}_{i}^{F_2 \setminus F_1}$, $i \in F_2 \setminus F_1$ are better than the corresponding $\hat{\mu}_{i}^{\text{MLE}}$ since they are larger and thus “closer to the truth”. For $i \in I \setminus F_2$, corresponding to the blue rows of Table 2, the components $\hat{\mu}_{i}^{F_2 \setminus F_1}$ are better than the $\hat{\mu}_{i}^{\text{MLE}}$ since, by construction, the $\hat{\mu}_{i}^{F_2 \setminus F_1}$ are infinite.

For reference, we list the estimates of the top five cell counts obtained using our method and compare them with those obtained by other methods in Dobra and Lenkoski (2011) in Appendix E.
6. Computing faces for large complexes. If our statistical model contains many variables and is not reducible, the problem of determining $F_t$ quickly becomes infeasible. Not only does the marginal polytope become very complicated, but also the size of the objects that one has to store or compute grows exponentially. Consider for example a $10 \times 10$ grid of binary random variables. This hierarchical model has 280 parameters, and the total sample space has cardinality $|I| = 2^{100} \approx 1.27 \times 10^{30}$. If $F_t$ is close to $I$, we cannot even list the elements of $F_t$, which consists of approximately $10^{30}$ elements. Therefore, we take a local approach and look for separators.

If the simplicial complex $\Delta$ contains a complete separator separating $V_1$ from $V_2$ in $\Delta$. We want to use $F_2 := F_{\Delta|V_1}(I_+) \cap F_{\Delta|V_2}(I_+)$ as an outer approximation and $F_1 := F_{\Delta_0}(I_+)$ as an inner approximation to $F_t$. Due to the problems mentioned above, we do not directly compute $F_1$ and $F_2$, but we compute their projections on $V_1$ and $V_2$. Instead of $F_2$, we compute the facial set $F_{2|V_1} := F_{\Delta|V_1}(\pi_{V_1}(I_+))$ of the $V_1$-marginal $\pi_{V_1}(I_+)$ with respect to $\Delta|V_1$. Similarly we compute $F_{2|V_2} := F_{\Delta|V_2}(\pi_{V_2}(I_+))$. Instead of $F_1$, we compute $F_{1|V_1} := F_{\Delta|V_1}(\pi_{V_1}(I_+))$ and $F_{1|V_2} := F_{\Delta|V_2}(\pi_{V_2}(I_+))$. Then we could recover $F_1$ and $F_2$ from the equations

$$F_2 = \pi_{V_1}^{-1}(F_{2|V_1}) \cap \pi_{V_2}^{-1}(F_{2|V_2}) \quad \text{and} \quad F_1 = \pi_{V_1}^{-1}(F_{1|V_1}) \cap \pi_{V_2}^{-1}(F_{1|V_2}).$$

For any $x \in I$, we can check whether $x \in F_1$ by checking whether $\pi_{V_1}(x) \in F_{1|V_1}$ and $\pi_{V_2}(x) \in F_{1|V_2}$. More importantly, we can check whether $F_1 = F_2$ by checking whether $F_{1|V_1} = F_{2|V_1}$ and $F_{1|V_2} = F_{2|V_2}$. This idea can be applied iteratively when either $\Delta|V_1$ or $\Delta|V_2$ has a separator.

The next two subsections illustrate these ideas. In Section 6.1, we consider a graph with no particular regularity pattern on 50 nodes and identify convenient separators. In Section 6.2, we consider a grid graph and work with two families of “parallel” separators that can be used to iteratively improve the inner approximation.

6.1. US Senate Voting Records Data. We consider the voting record of all 100 US senators on 309 bills from January 1 to November 19 2015. Similar
data for the years 2004–2006 was analyzed by Banerjee et al. (2008). The votes are recorded as “yea,” “nay” or “not voting.” We transformed the “not voting” into “nay” and consequently have a 100-dimensional binary data set. To fit a hierarchical model to this data set, we use the $\ell_1$-regularized logistic regression method proposed by Ravikumar et al. (2010) to identify the neighbours of each variable and construct an Ising model. We set the regularization parameter to $\lambda = 32 \sqrt{\log p / n} \approx 0.35$. The underlying graph of the Ising model is given in Figure 2. This figure should not be interpreted as the graph of a graphical model. Rather, the edges in the graph indicate where the two-way interactions lie. There are 277 parameters in this model (the number of vertices plus the number of edges). The graph consists of
two separate large connected components and 14 independent nodes.

There are 309 sample points, and \(|I_+| = 278\). We want to characterize the face \(F_t\) of the data on the marginal polytope. The graph in Figure 2 has many complete separators, and it decomposes as a union of several small irreducible simplicial subgraphs and two large irreducible subgraphs, one in each of the large connected components, as shown in Figure 3. By Lemma 2.5, we can restrict attention to these irreducible subgraphs. For the small irreducible subgraphs, one easily verifies that the data does not lie on a proper face of their corresponding marginal polytopes. We are left with the two large irreducible prime components in Figure 3.

**Table 3**

| ID | Senator | ID | Senator | ID | Senator | ID | Senator |
|----|---------|----|---------|----|---------|----|---------|
| 22 | Nelson  | 37 | Cardin  | 52 | Murphy  | 61 | Whitehouse |
| 23 | Reed    | 41 | Markey  | 53 | Hirono  | 87 | Warren   |
| 26 | Schumer | 47 | Udall   | 56 | Gillibrand |

*Numbering of some senators*

![Graph](image)

**Fig 3.** The simplicial complexes after cutting off the small prime components: (a) the Republican party prime component \(\Delta_r\), (b) the Democratic party prime component \(\Delta_d\). The yellow and pink nodes are the two separator sets we used to compute the facial set.

The Democratic party simplicial complex \(\Delta_d\) consists of 26 variables, and the model induced from \(\Delta_d\) contains 77 parameters, which is too large to use linear programming to compute the face of \(\mathbf{P}_{\Delta_d}\) containing the vector \(t_d\). Therefore, we look for separators in order to obtain good inner and outer approximations. Figure 3b indicates (in yellow and pink) two separators that separate \(\Delta_d\) into three simplicial complexes denoted, from left to right, by \(\Delta_\alpha\), \(\Delta_\beta\) and \(\Delta_\gamma\) that are small enough for the linear programming method. \(\Delta_\alpha\) has 9 nodes. The corresponding vector \(t_\alpha\) lies in the relative interior of \(\mathbf{P}_{\Delta_\alpha}\).

\(\Delta_\beta\) has 13 nodes, and \(t_\beta\) lies on a facet \(F_t\) of \(\mathbf{P}_{\Delta_\beta}\). To simplify the
notation, we denote the 100 senators not by their name but by an integer between 1 and 100. We only need to identify a few and their numbers are given in Table 3. The inequality of $F_{t \alpha}$ is

$$t_{87} - t_{56,87} \geq 0,$$

where $t_{87}$ denotes the marginal count of senator Warren voting “yea” and $t_{56,87}$ denotes the marginal counts of both senators Gillibrand and Warren voting “yea.”

$\Delta_\gamma$ has 11 nodes. The data vector $t_\gamma$ lies on the facet of $P_{\Delta_\gamma}$ with inequality

$$t_{23} - t_{23,53} \geq 0.$$

The intersection of the two facets (15) and (16) gives the outer approximation $F_{t \delta,d}$ to $F_{t \delta}$. To get an inner approximation, we complete each separator, i.e. the yellow vertices are completed and the pink vertices are completed in Figure 3b. Denote the three simplicial complexes with complete separators as $\Delta_\alpha$, $\Delta_\beta$, $\Delta_\gamma$ respectively. Then $\Delta_\delta = \Delta_\alpha \cup \Delta_\beta \cup \Delta_\gamma$ is a simplicial complex with two complete separators. The smallest face $F_{t \delta}$ of the marginal polytope $P_{\Delta_\delta}$ containing the data vector $t_\delta$ is our inner approximation. The models of $\Delta_\alpha$, $\Delta_\beta$, $\Delta_\gamma$ and $\Delta_\delta$ include main effects, two-, three- and four-way interactions. The dimension of the model induced by $\Delta_\delta$ is 91 (completing the two separators adds 14 parameters to the original model). We apply the linear programming method to $P_{\Delta_\alpha}$, $P_{\Delta_\beta}$ and $P_{\Delta_\gamma}$.

The dimension of the model of $\Delta_\alpha$ is 27, and $F_{t \alpha}$ is a facet with equation

$$\langle g_1, t_\alpha \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0.$$

It follows that $\{g_1\}$ is a basis of the kernel of $A_{F_\alpha}$.

The dimension of the model for $\Delta_\beta$ is 48. The face $F_{t \beta}$ has codimension 5, with defining equations

$$\begin{cases}
\langle g_2, t_\beta \rangle = t_{87} - t_{56,87} = 0 \\
\langle g_3, t_\beta \rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\
\langle g_4, t_\beta \rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\
\langle g_5, t_\beta \rangle = t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0 \\
\langle g_6, t_\beta \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0
\end{cases}$$

Again, $\{g_2, g_3, g_4, g_5, g_6\}$ is a basis of the kernel of $A_{F_\beta}$. 
The dimension of the model for $\Delta_\gamma$ is 38. The face $F_{t_\delta}$ has codimension 3. It is defined by the equations
\begin{equation}
\begin{align*}
\langle g_7, t_\gamma \rangle &= t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\
\langle g_8, t_\gamma \rangle &= t_{37,47,52} - t_{47,52,61} = 0 \\
\langle g_9, t_\gamma \rangle &= t_{23} - t_{23,53} = 0.
\end{align*}
\end{equation}
(19)

Again, $\{g_7, g_8, g_9\}$ is a basis of the kernel of $A_{F_\delta}$.

From Lemma 2.5, $F_{t_d} = F_{\tilde{\delta}} \cap F_{\tilde{\beta}} \cap F_{\tilde{\gamma}}$, and the equations for $F_{t_d}$ are
\begin{equation}
\begin{align*}
\langle g'_1, t_{\tilde{\delta}} \rangle &= t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0 \\
\langle g'_2, t_{\tilde{\delta}} \rangle &= t_{87} - t_{56,87} = 0 \\
\langle g'_3, t_{\tilde{\delta}} \rangle &= t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\
\langle g'_4, t_{\tilde{\delta}} \rangle &= t_{37,47,52,61} - t_{47,52,61} = 0 \\
\langle g'_5, t_{\tilde{\delta}} \rangle &= t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0 \\
\langle g'_6, t_{\tilde{\delta}} \rangle &= t_{23} - t_{23,53} = 0,
\end{align*}
\end{equation}
(20)
where the vectors $g'_1, \ldots, g'_9$ are the vectors $g_1, \ldots, g_9$ extended to $\mathbb{R}^{91}$ by adding zeros on the corresponding complementary coordinates (since $g'_1 = g'_6$, $g'_3 = g'_7$, $g'_4 = g'_8$, only six of the nine equations are needed). Thus, $F_{t_d}$, defined by (20), is a strict subset of the face $F_{t_d}$ defined by (15) and (16). Next, we refine our argument and show that indeed $F_{t_d} = F_{t_d}$.

From what we know, it follows that the orthogonal complement of the subspace generated by $F_{t_d}$ is
\[ G = \{ g' \in \mathbb{R}^{91} \mid g' = k_1 g'_1 + k_2 g'_2 + k_3 g'_3 + k_4 g'_4 + k_5 g'_5 + k_9 g'_9 \} \].

To describe $F_{t_d}$, we note that each defining equation of $F_{t_d}$ is of the form $\langle g, t_{\tilde{d}} \rangle = 0$, where $g$ is orthogonal to $F_{t_d}$. For any such $g$, let $g'$ be its extension to a vector in $\mathbb{R}^{91}$ by adding zero components. Then $g' \perp F_{t_d}$, which implies that $g' \in G$. Therefore, we can find $g$ by finding all vectors $g' \in G$ that vanish on all added components. This yields a system of linear equations in $k_1, \ldots, k_9, k_9$. We claim that all solutions must satisfy $k_1 = k_3 = k_4 = k_5 = 0$. Indeed, the coefficient of any triple or quadruple interaction must vanish (since these do not belong to the original Ising model), which implies $k_1 = k_3 = k_4 = 0$, and also the coefficient of $t_{37,52}$ must vanish, which implies $k_5 = 0$. On the other hand, the vectors $g'_2$ and $g'_9$ only contain interactions that are already present in $\Delta$, and so the coefficients $k_2$ and $k_9$ are free. Thus the equations for $F_{t_d}$ are
\begin{equation}
\begin{align*}
\langle g_2, t_{\tilde{\beta}} \rangle &= t_{87} - t_{56,87} = 0, \\
\langle g_9, t_{\tilde{\gamma}} \rangle &= t_{23} - t_{23,53} = 0.
\end{align*}
\end{equation}
(21)
This is the same as the outer approximation $F_{2,d}$.

The Republican simplicial complex $\Delta_r$ consists of 20 variables, and the model induced from $\Delta_r$ contains 46 parameters, which is also too large to directly compute $F_t$. The yellow nodes in Figure 3a separate $\Delta_r$ into two simplicial complexes denoted (from left to right) by $\Delta_a$ and $\Delta_b$. To compute the inner approximation, we complete the yellow separator and obtain two new simplicial complexes $\Delta_\tilde{a}$ and $\Delta_\tilde{b}$. With linear programming, we find that the corresponding vectors $t_\tilde{a}$ and $t_\tilde{b}$ lie in the relative interior of the polytopes $P_{\Delta_\tilde{a}}$ and $P_{\Delta_\tilde{b}}$, respectively. Therefore, $F_1 = P_{\Delta_r}$, from which we conclude that the corresponding vector $t_r$ lies in the relative interior of $P_{\Delta_r}$.

Thus, the face $F_t$ is now determined: it is characterized by the equalities (21). What insight is there in the knowledge (i) of the non-existence of the MLE and (ii) of the exact face $F_t$? While we have given general remarks in the introduction, let us illustrate here how knowledge about $F_t$ points to some issues with the statistical analysis that would possibly be overlooked if $F_t$ was not known.

First, knowing $F_t$, and its defining inequalities, for one model also gives information about other models. It follows from (21) that the MLE does not exist for any hierarchical model that includes one of the edges $(23,53)$ or $(56,87)$ (to see this, note that the inequality (16) defines a proper face for any model containing the edge $(23,53)$, since the corresponding sufficient statistics vector satisfies the equality in (16)). Thus, if one wants to find a smaller model, within the realm of hierarchical models, for which the MLE exists, both edges have to be dropped. However, from the data, here, evidence for both edges is quite strong, and thus the edges should not be dropped.

Second, let us consider the computation of the EMLE. As we know $F_t$, instead of running an MLE computation for a model with 277 parameters and $2^{100}$ outcomes, we are left with an MLE computation for a model with $277 - 2 = 275$ parameters and $|F_t| = 9 \cdot 2^{100}$ outcomes, those without the configurations $(X_{23}, X_{53}) = (1,0)$ or $(X_{56}, X_{87}) = (1,0)$ that all have counts zero. These numbers are still too large for a direct computation, even when taking into account that the EMLE can be computed by restricting to each of the irreducible components. So, we turn to an approximate method and compute the maximum composite likelihood estimate. The maximum composite likelihood estimate can be obtained by combining estimates from the local conditional likelihoods derived from the distribution of each variable given its neighbours: see for example Liu and Ihler (2012) or Massam and Wang (2017). Thus the reliability of the maximum composite likelihood estimates depends upon the existence of the maximum in each of the local
conditional likelihood. These local conditional likelihoods are derived from the global model built on the entire cell set $I$ and, certainly in practice, without worrying about the existence of the global MLE. Let us consider, for example, the likelihood obtained from the conditional distribution of $X_{23}$ given its neighbours $X_{19,53,61,78}$. For convenience, let $19,23,53,61,78$ be denoted as $a, b, c, d, e$. This likelihood is the product over all configurations of $i_{acde}$ in the data set of conditional binomial distributions for the variable $X_b$ and can be written as

$$
\prod_{i_a,i_b,i_c,i_d,i_e} p(X_b = i_b | X_{acde} = (i_a, i_c, i_d, i_e))^{n(i_a,i_b,i_c,i_d,i_e)},
$$

where $n(i_a,i_b,i_c,i_d,i_e)$ denotes the corresponding marginal cell count. It is easy to show that the MLE of each $p(X_b = 1 | i_a, i_c, i_d, i_e)$ is the empirical estimate $n(i_a,i_b,i_c,i_d,i_e)/n(i_a,i_c,i_d,i_e)$. In the data set, $n(i_a,i_b = 1, i_c = 0, i_d, i_e) = 0$ for all $i_a, i_d, i_e \in \{0,1\}$. Thus,

$$
\hat{p}(X_b = 1 | i_a = 1, i_c = 0, i_d = 1, i_e = 1) = \frac{\exp(\hat{\theta}_b + \hat{\theta}_{ab} + \hat{\theta}_{bd} + \hat{\theta}_{be})}{1 + \exp(\hat{\theta}_b + \hat{\theta}_{ab} + \hat{\theta}_{bd} + \hat{\theta}_{be})} = 0,
$$

so that $\hat{\theta}_b + \hat{\theta}_{ab} + \hat{\theta}_{bd} + \hat{\theta}_{be} = -\infty$, and the MLE of at least some of these parameters, which are the corresponding parameters of the global model, does not exist. Now the composite maximum likelihood estimate is obtained by averaging the estimates obtained from various local conditional likelihoods. From the $b$-local conditional model, we also obtain $\hat{p}(X_b = 1 | i_a = 1, i_c = 1, i_d = 1, i_e = 0) = 1/2$ and $\hat{p}(X_b = 1 | i_a = 0, i_c = 1, i_d = 0, i_e = 1) = 4/5$, which yield the linear combinations

$$
\hat{\theta}_b + \hat{\theta}_{bc} + \hat{\theta}_{ab} + \hat{\theta}_{bd} = 0, \quad \hat{\theta}_b + \hat{\theta}_{be} + \hat{\theta}_{bd} = 1.4.
$$

The remarks above are verified numerically. Let $\theta = (\theta_b, \theta_{ab}, \theta_{bc}, \theta_{bd}, \theta_{be})$, and denote by $l_{\text{local}}(\theta)$ the local conditional likelihood. Starting at $\theta_0 = (0,0,0,0,0)$ and optimizing (22) in terms of $\theta$ in Matlab, we obtain $l_{\text{local}}(\theta) = 3.88830675$ for the maximum $\hat{\theta} = \hat{\theta}_1 \approx (-62.3, -16.8, 35.2, 43.9, 28.5)$. If we change the starting point to $\theta_0 = (100,100,100,100,100)$, we obtain $\hat{\theta}_2 = (-162.2,-26.1,91.2,97.1,72.4)$ and $l_{\text{local}}(\theta_2) = 3.88830648$. Clearly, the values for $\theta$ are unreliable since the MLE in the local conditional model does not exist. However, both $\hat{\theta}_1$ and $\hat{\theta}_2$ satisfy equations (23). One can, of course, obtain estimates of $\theta_{bc}, \theta_{ab}, \theta_{bd}, \theta_{be}$ from the local conditional models centered at $c, a, d$ and $e$ respectively but these estimates will not have been obtained through the method of composite likelihood and it remains to study their properties.
This example shows that our method scales well and makes it possible to obtain the face $F_t$ for very large examples. It also illustrates how knowing the face $F_t$ gives us precious information on the reliability of the maximum composite likelihood estimate.

6.2. The $5 \times 10$-grid. Let $\Delta$ be the simplicial complex of the $5 \times 10$ grid graph. We exploit the regularity of this graph and make use of the vertical separators in the grid to obtain inner and outer approximations of the facial sets. The graph has 50 nodes, which is too many to directly compute a facial set or even to store it. However, the $5 \times 10$ grid has 8 vertical separators marked in red and blue in Figure 4, and we can use these to approximate $F_t$. Since facial sets for $5 \times 3$ grids can be computed reasonably fast (3 to 4 seconds on a laptop with 2.50 GHz processor and 12 GB memory), we only use three of these vertical separators at a time, say the blue separators $S_2 = \{11, \ldots, 15\}$, $S_4 = \{21, \ldots, 25\}$, $S_6 = \{31, \ldots, 35\}$, $S_8 = \{41, \ldots, 45\}$.

These separate the vertex sets

\[ V_1 = \{1, \ldots, 15\}, \quad V_3 = \{11, \ldots, 25\}, \quad V_5 = \{21, \ldots, 35\}, \]
\[ V_7 = \{31, \ldots, 45\}, \quad V_9 = \{41, \ldots, 50\}. \]

Adding the blue separators to $\Delta$ gives a simplicial complex

\[ \Delta_{S_2; S_4; S_6; S_8} := \Delta \bigcup_{j=2,4,6,8} \{ F : F \subseteq S_j \} \]

with five irreducible components supported on the vertex sets $V_1, V_3, V_5, V_7$ and $V_9$ (Figure 6). To compute a facial set with respect to $\Delta_{S_2; S_4; S_6; S_8}$,
according to Lemma 2.5 applied four times, we need to compute
\[ G_{1,V_1} := F_{\Delta_{S_2 \setminus V_1}}(\pi_{V_1}(I_+)), \quad G_{1,V_3} := F_{\Delta_{S_2 \setminus S_4 \setminus V_1}}(\pi_{V_3}(I_+)), \]
\[ G_{1,V_5} := F_{\Delta_{S_4 \setminus S_6 \setminus V_1}}(\pi_{V_5}(I_+)), \quad G_{1,V_7} := F_{\Delta_{S_6 \setminus S_8 \setminus V_1}}(\pi_{V_7}(I_+)), \]
\[ G_{1,V_9} := F_{\Delta_{S_8 \setminus V_1}}(\pi_{V_9}(I_+)). \]

Then \( G_1 := \bigcap_i \pi_{V_i}^{-1}(G_{1,V_i}) \) is equal to \( F_{\Delta_{S_2 \setminus S_4 \setminus S_6 \setminus S_8}}(I_+) \), and thus an inner approximation of \( F_\ell \). As stated before, we do not need to compute \( G_1 \) explicitly, but we represent it by means of the \( G_{1,V_i} \).

We can improve the approximations by also considering the red separators
\( S_1 = \{6, \ldots, 10\}, \quad S_3 = \{16, \ldots, 20\}, \quad S_5 = \{26, \ldots, 30\}, \quad S_7 = \{36, \ldots, 40\}, \)
that separate
\[ V_0 = \{1, \ldots, 10\}, \quad V_2 = \{6, \ldots, 20\}, \quad V_4 = \{16, \ldots, 30\}, \quad V_6 = \{26, \ldots, 40\}, \quad V_8 = \{36, \ldots, 50\}. \]

As explained in Section 3.1, we want to compute \( G_{1,V_1}^{(2)} := F_{\Delta_{S_1 \setminus S_3 \setminus S_5 \setminus S_7}}(G_1) \).

Again, instead of computing \( G_{1,V_1}^{(2)} \) directly, we need only compute the much smaller sets
\( G_{1,V_0}^{(2)} := \pi_{V_0}(G_1^{(2)}), \quad G_{1,V_2}^{(2)} := \pi_{V_2}(G_1^{(2)}), \ldots, \quad G_{1,V_9}^{(2)} := \pi_{V_9}(G_1^{(2)}). \)

So the question is: is it possible to compute \( G_{1,V_0}^{(2)}, \quad G_{1,V_2}^{(2)}, \quad \ldots, \quad G_{1,V_8}^{(2)} \) from \( G_{1,V_1}, \quad G_{1,V_3}, \quad \ldots, \quad G_{1,V_9} \), without computing \( G_1 \) in between?

It turns out that this is indeed possible: By Lemma 2.5, all we need to compute \( G_{1,V_i}^{(2)} \) is \( \pi_{V_j}(G_1), \quad j = i - 1, i + 1 \). For \( i = 0 \), since \( V_0 \subset V_1 \), we can compute \( G_{1,V_0} \) from \( \pi_{V_1}(G_1) = G_{1,V_1} \). For \( i = 2, 4, 6, 8 \), since \( V_i \subset V_{i-1} \cup V_{i+1} \), we can compute \( G_{1,V_i} \) from \( \pi_{V_{i-1} \cup V_{i+1}}(G_1) \), which itself can be obtained by “gluing” \( \pi_{V_{i-1}}(G_1) = G_{1,V_{i-1}} \) and \( \pi_{V_{i+1}}(G_1) = G_{1,V_{i+1}} \):
\[ \pi_{V_{i-1} \cup V_{i+1}}(G_1) = \left( \pi_{V_{i-1} \cup V_{i+1}}^{-1}(G_{1,V_{i-1}}) \right)^{-1} \left( \pi_{V_{i-1} \cup V_{i+1}}^{-1}(G_{1,V_{i+1}}) \right)^{-1}. \]
where $\pi_{V''}$ for $V'' \subseteq V'$ denotes the marginalization map from $I_{V'}$ to $I_{V''}$ and where $\left(\pi_{V''}\right)^{-1}$ denotes the lifting from $I_{V''}$ to $I_{V'}$.

As explained in Section 3.1, we have to iterate this procedure: From $G_1^{(2)}$ we want to compute $G_1^{(3)} := F_{\Delta_s2,s_4,s_6,s_8}(G_1')$ or, more precisely, we want to compute $G_{1,V_i}^{(3)} = \pi_{V_i}(G_1^{(3)})$ for $i = 1, 3, \ldots, 9$. Again, we do this without looking at $G_1^{(2)}$ directly by just using the information available through the $G_1^{(3)}$. Iterating this procedure, we obtain a sequence of sets $G_{1,V_i}^{(k)}$, $G_{1,V_j}^{(k)}$ (with odd $i$ and even $j$), which stabilizes after a finite number of steps. Let

$$F_{1,V_i} := \bigcup G_{1,V_i}^{(k)};$$

Our best inner approximation is then $F_1 = \bigcap_{i=0}^9 \pi_{V_i}^{-1}(F_{1,V_i})$. Again, we do not compute $F_1$ explicitly, but we represent it in terms of the $F_{1,V_i}$. The process is visualized in Figure 7.

Let us now consider the outer approximation $F_2$. We adapt Strategy 3 of Section 3.2 and cover the graph with $5 \times 3$ grid subgraphs, since the facial sets for such graphs can easily be computed. These subgrids are supported...
Table 4
facial set approximation of 5 × 10 grid graph

| sample size | $F_2 \neq I$ | $F_1 = F_2$ |
|-------------|---------------|---------------|
| 50          | 100.0%        | 94.3%         |
| 100         | 100.0%        | 82.5%         |
| 150         | 99.9%         | 76.5%         |
| 200         | 99.6%         | 81.2%         |
| 300         | 96.4%         | 87.7%         |
| 400         | 92.9%         | 91.5%         |
| 500         | 84.8%         | 93.9%         |
| 1000        | 44.7%         | 99.9%         |

We generated random data of varying sample size. For each fixed sample size, we generated 100 data samples. The simulation results are shown in Table 4. For each simulated sample, we compute the sets $F_{1,V_i}$ and $F_{2,V_i}$ as described above. When computing $F_{1,V_i}$, we found that 2 iterations actually suffice. Then we checked whether $F_2$ is a proper subset of $I$ (second column), and we checked whether $F_1 = F_2$ (third column). Both for small and large sample sizes, we found that the $F_1 = F_2$ quite often.

We also investigated what happens when the outer approximation is not computed using all $3 \times 5$-subgrids, but only a cover of four $3 \times 5$-subgrids and one $2 \times 5$-subgrid (as in Figure 5). In all simulations, this easier approximation gave the same result. The same is not true for the inner approximation: when using just one of the two families of parallel separators we obtain an inner approximation that is much too small.

7. Conclusion. As mentioned before, previous work had made it possible to identify $F_t$ for hierarchical models with up to 16 variables. In this paper, we offer a methodology to approximate, and sometimes, completely identify the facial set $F_t$ for high-dimensional models. To find an inner and an outer approximation to $F_t$, first, we divided the original problem into subproblems of dimension at most 16 for which we could use linear programming and, second, we combined the facial sets of the subproblems and related then to $F_t$. Identifying the subproblems and relating the facial sets to $F_t$ is numerically easy and the corresponding software can be obtained.
data $I_+$ on $5 \times 10$ grid

\[ I_+^1 = \pi_{V_1}(I_+) \]
\[ I_+^3 = \pi_{V_3}(I_+) \]
\[ I_+^5 := \pi_{V_5}(I_+) \]
\[ I_+^7 = \pi_{V_7}(I_+) \]
\[ I_+^9 = \pi_{V_9}(I_+) \]

FIG 7. Flow chart
It has long been established that determining the existence of the MLE is essential to correct statistical inference. In our paper, we have emphasized the problem of parameter estimation and shown how working with the likelihood $l_{F_2}$ yields much better estimates of the parameter that when working with $l$. When testing one model versus another, the correct degrees of freedom for the asymptotic distribution of the test statistic is the difference between the dimensions of the facial sets for the two models being compared and not the difference between the dimensions of the two models. If we only know approximations $F_1$ and $F_2$, we can use their dimensions to approximate the correct degrees of freedom.

In high dimensions, when the likelihood functions has so many terms that the (E)MLE cannot be computed, a popular approach is to compute the maximum composite likelihood estimate. We have shown through an example that, when the global MLE does not exist, the local MLE for some of the same parameters might not exist either. So, combining the values of the MLE of local likelihoods without being aware that the data lies on a face of the marginal polytope, one might also obtain misleading estimates of the parameters through composite likelihood.

We have not addressed the question of how to obtain reliable confidence intervals for the parameters by exploiting the properties of the inner and outer approximations to $F_t$. This subject clearly deserves attention and should be the subject of further work.

**APPENDIX A: PARAMETRIZING HIERARCHICAL MODELS**

In this section, we recall the usual parametrization of hierarchical models, see, for example, Letac and Massam (2012). The starting point is the parametrization (2) of the hierarchical model, which we repeat here for convenience:

\[
\text{(A24)} \quad \log p(i) = \sum_{D \in \Delta} \theta_D(i_D)
\]

This parametrization is not identifiable; that is, for any joint distribution $p$ from the hierarchical model there are different choices for the functions $\theta_D$ that satisfy (A24). One way to make the parameters unique is to choose a special element within each set $I_v$, which we denote by 0. The choice of 0 is arbitrary, and a different choice of 0 leads to a simple affine change of parameters. With this choice, the functions $\theta_D$ become unique if one requires $\theta_D(i_D) = 0$ whenever $i_v = 0$ for some $v \in D$. 

A parametrization in terms of real numbers is obtained using the following definitions: for \( i \in I \), we write

\[
S(i) = \{ v \in V ; \ i_v \neq 0 \}, \quad J = \{ j \in I \setminus \{ 0 \}, \ S(j) \in \Delta \}.
\]

For any \( j \in J \), let

\[
\theta_j = \theta_D(i_D) \quad \text{for the unique} \quad i \in I \quad \text{with} \quad S(i) = D, \ i_D = j_D.
\]

To simplify the notation, we write \( j \triangleleft i \) whenever \( S(j) \subseteq S(i) \) and \( j \in S(i) \). It is convenient to introduce the vectors

\[
f_i = \sum_{j \in J : j \triangleleft i} e_j, \quad i \in I
\]

where \( e_j, j \in J \) are the unit vectors in \( \mathbb{R}^J \). 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 \( \left( \frac{1}{f_i} \right), \ i \in I \). Then (2) can be rewritten in the following equivalent forms

\[
(A25) \quad \log p_\theta(i) = \sum_{j \in J : j \triangleleft i} \theta_j - k(\theta) = \langle \theta, f_i \rangle - k(\theta) = A^t \theta - k(\theta) = \tilde{A}^t \tilde{\theta},
\]

where \( \tilde{\theta} = (\theta_0, \theta) \) as a column vector and

\[
(A26) \quad -\theta_0 = k(\theta) = \log \left( \sum_{i \in I} \exp \left( \sum_{j \in J : j \triangleleft i} \theta_j \right) \right)
\]

acts as a normalization constant. If \( n = (n(i), i \in I) \) denotes the \( I \)-dimensional column vector of cell counts, then

\[
(A27) \quad \tilde{A} n = \begin{pmatrix} N \\ t \end{pmatrix} \quad \text{and} \quad An = t,
\]

where \( N = \sum_{i \in I} n(i) \) is the total cell counts and \( t \) is the column vector of sufficient statistic with components equal to the \( j_{S(j)} \)-marginal counts \( n(j_{S(j)}) \), i.e. \( t = (t_j, j \in J) \) where \( t_j = n(j_{S(j)}) = \sum_{i \in I : j_{S(j)} = j} n(i), \ j \in J \).

It follows from (A27) that \( \frac{1}{N} = \sum_{i \in I} \frac{n(i)}{N} f_i \). Therefore, \( t \) belongs to the convex polytope with extreme points \( f_i, i \in I \). This polytope is the marginal polytope of the hierarchical model, denoted by \( P_\Delta \).
Example A.1. Let $V = \{a, b, c\}$, $I_a = \{0, 1\} = I_b = I_c$ and $\Delta = \{a, b, c, ab, bc\}$. Then

\[ I = \{000, 100, 010, 110, 001, 101, 011, 111\}, \]

\[ J = \{(100), (010), (001), (110), (011)\}, \]

\[ \tilde{A} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \theta_{000} \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \theta_{100} \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & \theta_{010} \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \theta_{001} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & \theta_{110} \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \theta_{011}
\end{pmatrix} \]

APPENDIX B: EXAMPLE: TWO BINARY RANDOM VARIABLES

Consider two binary random variables, and let $\Delta = \emptyset, \{1\}, \{2\}, \{1, 2\}$.

The hierarchical model $E_\Delta$ is the saturated model; that is, it contains all possible probability distributions with full support. Then

\[ \tilde{A} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & \theta_{000} \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & \theta_{010} \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & \theta_{100} \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & \theta_{001} \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & \theta_{110} \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & \theta_{011}
\end{pmatrix} \]

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

\[ F_{00} : 1 - t_{01} - t_{10} + t_{11} \geq 0, \quad F_{01} : t_{01} - t_{11} \geq 0, \quad F_{10} : t_{10} - t_{11} \geq 0, \quad F_{11} : t_{11} \geq 0. \]

Each of the corresponding facets contains three columns of $A$. In fact, the facet $F_i$ 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 $F_{00}$ (i.e. $n = (0, n_{01}, n_{10}, n_{11})$ with $n_{01}, n_{10}, n_{11} > 0$). If $p_0(t) \to p_*$, then $p_0(00) \to 0$, while all
other probabilities converge to a non-zero value. It follows that

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

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 \(\theta_{ij} := \lim_{s \to \infty} \theta_{ij}^{(s)}:\)

| facet | \(\theta_{00}\) | \(\theta_{01}\) | \(\theta_{10}\) | \(\theta_{11}\) | finite parameter combinations: |
|-------|----------|----------|----------|----------|----------------------------------|
| \(F_{00}\) | \(-\infty\) | \(+\infty\) | \(+\infty\) | \(-\infty\) | \(\theta_{01}^{(s)} + \theta_{00}^{(s)} + \theta_{00}^{(s)} + \theta_{10}^{(s)} + \theta_{10}^{(s)} + \theta_{11}^{(s)} + \theta_{11}^{(s)}\) |
| \(F_{01}\) | finite | \(-\infty\) | finite | \(+\infty\) | \(\theta_{00}^{(s)}, \theta_{10}^{(s)}, \theta_{01}^{(s)} + \theta_{11}^{(s)}\) |
| \(F_{10}\) | finite | finite | \(-\infty\) | \(+\infty\) | \(\theta_{00}^{(s)}, \theta_{01}^{(s)}, \theta_{10}^{(s)} + \theta_{11}^{(s)}\) |
| \(F_{11}\) | 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 \(\theta_{ij}^{(s)}\) 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 \(\mu_i\) introduced in Section 4.2. For example, on the facet \(F_{01}\), consider the parameters

\[
\begin{align*}
\mu_{10} &= \log p(10)/p(00) = \theta_{10}, \\
\mu_{11} &= \log p(11)/p(00) = \theta_{10} + \theta_{01} + \theta_{11}, \\
\mu_{01} &= \log p(01)/p(00) = \theta_{01}.
\end{align*}
\]

Then \(\mu_{10}\) and \(\mu_{11}\) are identifiable parameters on \(E_{F_{01}}\), and \(\mu_{01}\) diverges close to \(F_{01}\). 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 \(\theta_{00}\), which is not a linear combination of those but that is fine because it is not free.

We obtain similar results for the facets \(F_{01}\) and \(F_{11}\). The results are summarized in the following table:

| facet | \(\mu_{01}\) | \(\mu_{10}\) | \(\mu_{11}\) |
|-------|----------|----------|----------|
| \(F_{01}\) | \(-\infty\) | finite | finite |
| \(F_{10}\) | finite | \(-\infty\) | finite |
| \(F_{11}\) | finite | finite | \(-\infty\) |
Of course, by definition of the $\mu_i$s, we cannot consider the facet $F_{00}$ where $n_i(00) = 0$. To study $F_{00}$, we have to choose another zero cell and redefine the parameters $\mu_i$.

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 $E_{\Delta,F}$. Thus, in this case there are more remaining parameters than the dimension of $E_{\Delta,F}$.

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

$$
\theta^{(s)}_{00} = \log p_{\theta(s)}(00) \to \frac{n_{00}}{N},
$$

$$
\theta^{(s)}_{01} = \log \frac{p_{\theta(s)}(01)}{p_{\theta(s)}(00)} \to -\infty,
$$

$$
\theta^{(s)}_{10} = \log \frac{p_{\theta(s)}(10)}{p_{\theta(s)}(00)} \to \log \frac{n_{10}}{n_{00}}.
$$

However, the limit of $\theta^{(s)}_{11} = \log \frac{p_{\theta(s)}(11)p_{\theta(s)}(00)}{p_{\theta(s)}(01)p_{\theta(s)}(10)}$ is not determined. The only constraint is that $\theta^{(s)}_{11}$ cannot go to $+\infty$ faster than $\theta^{(s)}_{01}$ goes to $-\infty$, since $p_{\theta(s)}^{(1)} = \exp(\theta^{(s)}_{00} + \theta^{(s)}_{01} + \theta^{(s)}_{10} + \theta^{(s)}_{11})$ 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 $\theta_j$ in turn. To be precise, we order the parameters $\theta_j$ in some way. For simplicity, say that the parameters are $\theta_1, \theta_2, \ldots, \theta_h$. Then we let

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

(this is called the non-linear Gauss-Seidel method). Let us choose the ordering $\theta_0, \theta_1, \theta_{11}$ (note that $\theta_0 = -k(\theta)$ is not a free parameter). We start at $\theta_0^{(0)} = \theta_1^{(0)} = \theta_{11}^{(0)} = 0$. In the first step, we only look at $\theta_0$. That is, we want to solve

$$
0 = \frac{\partial}{\partial \theta_{01}} l(\theta) = -\frac{\exp(\theta_0^{(1)}) + \exp(\theta_1^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_0^{(1)}) + \exp(\theta_1^{(0)}) + \exp(\theta_0^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})} = -\frac{2\exp(\theta_0^{(1)})}{1 + 2\exp(\theta_0^{(1)})}.
$$

This derivative is negative for any finite value of $\theta_0^{(1)}$, 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 $\theta_{10}$. We fix the other variables and try to solve
\[
0 = \frac{\partial}{\partial \theta_{10}} l(\theta) = \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{11}^{(1)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}
\]
\[
\approx \frac{n_{10}}{N} - \frac{\exp(\theta_{10}^{(1)})}{1 + \exp(\theta_{10}^{(1)})},
\]
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 $\theta_{11}$. 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 $\theta_{11}$. 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 $\theta_{01}$ 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 $\theta_{11}$ is optimized before $\theta_{01}$, then $\theta_{11}^{(1)}$ will in any case be a large negative number.

For general data, the derivative of $l(\theta)$ with respect to $\theta_{01}$ (equation (B28)) takes the form
\[
\frac{\partial}{\partial \theta_{01}} l(\theta) = \frac{t_{01}}{N} - \frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(1)} + \theta_{11}^{(1)})}.
\]
Setting this derivative to zero and solving for $\theta_{01}^{(1)}$ leads to a linear equation in $\theta_{01}^{(1)}$ with symbolic solution
\[
\theta_{01}^{(1)} = \log \frac{1 + \exp(\theta_{10}^{(0)})}{1 + \exp(\theta_{10}^{(0)} + \theta_{11}^{(0)})} \frac{t_{01}}{N} - \frac{t_{01}}{N}.
\]
In fact, for any hierarchical model, the likelihood equation is linear in any single parameter $\theta_j$, 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 (unfortunately, IPFP does not always fail that quickly when the MLE does not exist).

**APPENDIX C: PARAMETRIZATIONS ADAPTED TO FACIAL SETS**

Let us briefly discuss how to remedy problems 1. (identifiability), 2. (relation between parameters on $E$ and $E_{F_t}$) and 3. (cancellation of infinities in linear combinations of diverging parameters) from the beginning of Section 4.2. The idea to remedy 1. and 2. is to define parameters $\mu_i$, $i \in L$, of $E_A$ such that a subset $L_t \subseteq L$ of the parameters parametrizes $E_{F_t}$, in a consistent way. Denote by $A_{\mu} = (a_{\mu,j,i}, j \in L, i \in I)$ the design matrix of $E_A$ corresponding to the new parameters $\mu$. Then the necessary conditions are:

(*) Let $A_{L_t,F_t}^\mu := (a_{\mu,j,i}, j \in L_t, i \in F_t)$ be the submatrix of $A^\mu$ with rows indexed by $L_t$ and columns indexed by $L_t$, and denote by $\tilde{A}_{L_t,F_t}^\mu$ the same matrix with an additional row of ones. The rank of $\tilde{A}_{L_t,F_t}^\mu$ is equal to $|L_t| + 1$, the number of its rows (and thus, $A_{L_t,F_t}^\mu$ has rank $|L_t|$).

(**) $a_{\mu, j, i} = 0$ for all $i \in F_t$ and $j \in L \setminus L_t$.

In fact, (***) implies that $A_{L_t,F_t}^\mu$ is the design matrix of $E_{A,F_t}$, since the parameters $\mu_i$ with $i \notin L_t$ do not play a role in the parametrization $\mu \mapsto p_{F_t,\mu}$. Moreover, (*) implies that the parametrization $\mu \mapsto p_{F_t,\mu}$ is identifiable. In this sense, we have remedied problem 1.

Since $\tilde{A}_{L_t,F_t}^\mu$ has full row rank, it has a right inverse matrix $\tilde{C}$, such that $\tilde{A}_{L_t,F_t}^\mu \tilde{C} = I_{|L_t|+1}$ equals the identity matrix of size $|L_t| + 1$. Recall that

$$
\log p_{F_t,\mu}(i) = \langle \mu^I, f_i^\mu \rangle - k_F(\mu),
$$

$$
\log p_{\mu}(i) = \langle \tilde{\mu}^I, f_i^\mu \rangle - k(\mu),
$$

for any parameter vector $\mu$ and all $i \in F_t$. Since $f_i^\mu$ are the columns of $A^\mu$ and since the components of $f_i^\mu$ corresponding to $L \setminus L_t$ vanish by (**), we may apply the matrix $C$ obtained from $\tilde{C}$ by dropping the row corresponding to $k_F$ or $k$ and obtain

(C29) \hspace{1cm} (\log p_{\mu})C = \mu_{L_t} \quad \text{and} \quad (\log p_{F_t,\mu})C = \mu_{L}.
When \( p_{\mu(s)} \) is a sequence in \( \mathcal{E}_A \) with limit \( p_\mu \) in \( \mathcal{E}_{F_t,A} \), then (C29) shows that \( \mu_i^{(s)} \to \mu_i \) for \( i \in L_t \). In this sense, we have remedied problem 2.

Finally, we solve problem 3. Suppose that we have chosen parameters \( \mu_L \) as in Section 4.2, and let \( A^{\mu_L} \) be the design matrix with respect to these parameters. Then \( (A^{\mu_L})_{j,i} = 0 \) if \( i \in F_t \) and \( j \notin L_t \). Moreover, for \( j \in L_t \), the \( j \)th column of \( A_{\mu_L} \) has a single non-vanishing entry (equal to one) at position \( j \). Suppose that \( F_t \) corresponds to a face \( F_t \) of codimension \( c \). Then there are \( c \) facets of \( P \) whose intersection is \( F_t \). Thus, following the notation introduced in Remark 2.2, there exist \( c \) inequalities

\[
\langle \tilde{g}_1, \tilde{x} \rangle \geq 0, \quad \ldots, \quad \langle \tilde{g}_c, \tilde{x} \rangle \geq 0
\]

that together define \( F_t \). In this case, the vectors \( \tilde{g}_1, \ldots, \tilde{g}_c \) are linearly independent and satisfy \( \langle \tilde{g}_j, \tilde{f}_i \rangle = 0 \) (thus, they are a basis of the kernel of \( (A^{\mu_L}_{F_t})^t \)). It follows that the \( k \)th component of \( g_j \), denoted by \( g_{j,k} \), vanishes if \( k \in L_t \); that is, the inequalities (C30) do not involve the variables corresponding to \( L_t \). Let \( G \) be the square matrix, indexed by \( L \setminus L_t \) with entries \( g_{j,k} \), \( j, k \in L \setminus L_t \). Then the square matrix

\[
\tilde{G} = \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix}
\]

is invertible. We claim that the parameters \( \lambda = \tilde{G}^{-1} \mu_L \) are what we are looking for.

The design matrix with respect to the parameters \( \lambda \) is \( A^\lambda = \tilde{G} A^{\mu_L} \). For any \( j \notin L_t \),

\[
A^\lambda_{j,i} = 0, \quad \text{if} \quad i \in F_t, \quad \text{and} \quad A^\lambda_{j,i} = \langle \tilde{g}_j, \tilde{f}_i \rangle \geq 0, \quad \text{if} \quad i \notin F_t.
\]

This implies the following properties:

1. If all parameters \( \lambda_j \) with \( j \notin L_t \) are sent to \(-\infty\), then \( p_\lambda \) tends towards a limit distribution with support \( F_t \).

2. The coefficient of \( \lambda_j \) in any log-probability is non-negative, so there is no cancellation of \( \pm \infty \).

So far, we only used the fact that the vectors \( \tilde{g}_j \) define valid inequalities for the face \( F_t \). Suppose that we choose \( \tilde{g}_j \) in such a way that each inequality \( \langle \tilde{g}_j, \tilde{x} \rangle \geq 0 \) defines a facet. The intersection of less than \( c \) facets is a face that strictly contains \( F_t \). This implies that for each \( j \), there exists \( i_j \in I \setminus F_t \) such that \( f_{i_j} \) satisfies

\[
\langle \tilde{g}_j, \tilde{f}_{i_j} \rangle > 0, \quad \text{and} \quad \langle \tilde{g}_{j'}, \tilde{f}_{i_j} \rangle = 0 \quad \text{for all} \quad j' \neq j,
\]
and so

\[ A_{j, i_j}^\lambda > 0, \quad \text{and} \quad A_{j', i_j}^\lambda = 0 \text{ for all } j' \neq j. \]

Hence:

3. If \( \lambda_j^{(s)} \) are sequences of parameters such that \( p_{\lambda_j^{(s)}} \) tends towards a limit distribution with support \( F_t \), then \( \lambda_j^{(s)} \rightarrow -\infty \) for all \( j \notin L_t \).

It is not difficult to see that, conversely, any parametrization that satisfies these three properties comes from facets defining the face \( F_t \).

APPENDIX D: UNIFORM SAMPLING FOR THE 4 × 4-GRID

This section enhances the example in Section 5.1. In a second experiment, we generated random samples from the uniform distribution, that is from the probability distribution \( P_\theta \) in the hierarchical model where all parameters \( \theta_j, j \in J \), are set to zero. For each sample size, 1000 samples were obtained. The results are given in the following table:

| sample size | MLE does not exist | \( F_1 = F_t \) | \( F_2 = F_t \) |
|-------------|--------------------|----------------|----------------|
| 10          | 98.5%              | 96.3%          | 100.0%         |
| 15          | 68.9%              | 99.9%          | 100.0%         |
| 20          | 29.0%              | 100.0%         | 100.0%         |
| 50          | 0.0%               | 100.0%         | 100.0%         |

As the table shows, for larger samples the probability that a random sample lies on a proper face becomes very small. If \( F_t = I \), then clearly \( F_t = F_2 \). But we also found \( F_1 = F_2 \) for all samples with \( t \) lying on a proper face, which shows that \( F_2 \) is an excellent approximation of \( F_t \) in this model. For the inner approximation, we observed some samples with \( F_1 \neq F_t \), but they seem to be very rare.

APPENDIX E: ESTIMATED CELL FREQUENCIES FOR THE NLTCS DATA

The following table lists the estimates of the top five cell counts obtained using our method and compares them with those obtained by other methods in Dobra and Lenkoski (2011).
APPENDIX F: A LINEAR PROGRAMMING ALGORITHM TO COMPUTE FACIAL SETS

Let $A$ be the design matrix, let $A_+$ be the sub-matrix with columns indexed by the positive cells $I_+$, and let $A_0$ be the sub-matrix indexed by the empty cells (thus $A = (A_+, A_0)$ after reordering the columns).

**Lemma F.1.** Any solution $g^*$ of the non-linear problem

$$
\max \|Ag\|_0 \quad \text{subject to} \begin{cases} 
A_+g = 0, \\
A_0g \geq 0,
\end{cases}
$$

is perpendicular to the smallest face containing $t$. The corresponding facial set is $F_t = I \setminus \text{supp}(Ag^*)$.

The optimization problem (F31) is highly non-linear and non-convex. It can be solved by repeatedly solving the associated $\ell_1$-norm optimization problem:

$$
\max \|A_0g\|_1 \quad \text{subject to} \begin{cases} 
A_+g = 0, \\
A_0g \geq 0, \\
A_0g \leq 1.
\end{cases}
$$

Due to the constraints, $\|A_0g\|_1 = A_0g$, and so Problem (F32) is a linear programming problem. It can be solved repeatedly until the smallest facial set $F_t$ is obtained. The process is as follows:

The algorithm is introduced in the supplementary material to (Fienberg and Rinaldo, 2012), where it is also proved that it outputs the correct result.

APPENDIX G: SOME PROOFS

**G.1. Proof of Theorem 2.3.** Theorem 2.3 goes back to Barndorff-Nielsen (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:
**Algorithm 1** Computing \( F_t \) Linear Programming

**Require:** Design matrix \( A \) and positive cell index \( I_+ \)

Let \( A_+ := A(I_+, :) \), \( A_0 := A \setminus A_+ \)

repeat
    Solve problem (F32), get the solution \( g^* \) and the corresponding maximum \( z^* \)
    \( A_0 := \) the submatrix of \( A_0 \) with those columns \( f_i \) of \( A_0 \) that satisfy \( \langle f_i, g^* \rangle = 0 \).
until \( A_0 = \emptyset \) or \( z^* \neq 0 \)
if \( A_0 = \emptyset \) then
    \( F_t := I_+ \)
end if
if \( z^* = 0 \) then
    \( F_t := I_+ \cup \{ i \mid \text{column } i \text{ of } A \text{ belongs to } A_0 \} \)
end if

**Lemma G.1.** Let \( p \in \overline{E_A} \). Then \( p \in \mathcal{E}_{A,\text{supp}(p)} \).

**Lemma G.2.** Let \( p \in \overline{E_A} \). Then \( \mathcal{E}_{A,\text{supp}(p)} \subseteq \overline{E_A} \).

**Lemma G.3.** Let \( p \in \overline{E_A} \). Then \( \text{supp}(p) \) is facial.

**Lemma G.4.** If \( F \) is facial, then there exists \( p \in \overline{E_A} \) with \( \text{supp}(p) = F \).

Indeed, Lemma G.1 shows that \( \overline{E_A} \subseteq \bigcup_F \mathcal{E}_{A,F} \), where the union is over all support sets \( F \). Lemma G.2 shows the converse containment, so that \( \overline{E_A} = \bigcup_F \mathcal{E}_{A,F} \). 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 G.3 and G.4.

In the proofs of Lemmas G.1 to G.4, we need the following easy lemma of which we omit the proof:

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

**Proof of Lemma G.1.** Let \( p = \lim_{k \to \infty} p_k \), where \( p_k \in \mathcal{E}_A \), and let \( F = \text{supp}(p) \). 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 G.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 \( p \in \mathcal{E}_{A,F} \). \( \square \)

**Proof of Lemma G.2.** Let \( p = \lim_{k \to \infty} p_k \), where \( p_k \in \mathcal{E}_A \), let \( F = \text{supp}(p) \), and let \( q \in \mathcal{E}_{A,F} \). Then there exists parameters \( \theta \) with \( \log(q(i)) \) \( \log(p(i)) = (\theta, f_i) \) for all \( i \in F \). For any \( k \), there exists \( c_k > 0 \) such that \( q_k := c_k p_k \exp((\theta, A)) \in \mathcal{E}_A \). Then \( q_k \to q \) as \( k \to \infty \), and so \( q \in \overline{E_A} \). \( \square \)
Proof of Lemma G.3. Let \( p = \lim_{k \to \infty} p_k \), where \( p_k \in \mathcal{E}_A \), and let \( F = \mathcal{F}_A(\text{supp}(p)) \). Then \( x = \frac{1}{|\text{supp}(p)|} \sum_{i \in \text{supp}(p)} f_i \) 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

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

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

\[
\sum_{i \in F \setminus \text{supp}(p)} \lambda_i \log(p_k(i)) = \sum_{i \in \text{supp}(p)} \log(p_k(i))v_i \to \sum_{i \in \text{supp}(p)} \log(p(i))v_i.
\]

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

Proof of Lemma G.4. If \( F \) is facial, there exist \( g \in \mathbb{R}^h \) and \( c \in \mathbb{R} \) 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 \( F \). \( \square \)

G.2. Proof of Theorem 2.4. By definition, any EMLE \( p_* \) belongs to the closure of the model. According to Theorem 2.3, the support of \( p_* \) is facial. If \( \text{supp}(p) \) does not contain \( \text{supp}(n) \), then the log-likelihood goes to minus infinity, \( \hat{l}(p) = -\infty \), and so \( p \) does not maximize the likelihood. Therefore, \( \text{supp}(p_*) \) is a facial set containing \( \text{supp}(n) \). Thus, \( F_t \subseteq \text{supp}(p_*) \).

By Lemma G.1, \( p_* \) belongs to \( \mathcal{E}_{\Delta, \text{supp}(p_*)} \), which is parametrized by a vector \( \theta \), see Theorem 2.3. On \( \mathcal{E}_{\Delta, \text{supp}(p_*)} \), the log-likelihood function in terms of this parameter \( \theta \) is

\[
l_F(\theta) = \sum_{j \in J} \theta_j t_j - Nk_F(\theta).
\]
The critical equations satisfied by $p_\ast$ are

$$Ap_\ast = \frac{t}{N},$$

proving the first property. Note that these equations are independent of the parameters and the support of $p_\ast$. We now show that any solution to these equations is supported on the same face of $P$ as $\frac{t}{N}$.

Let $p$ be a probability distribution on $I$ such that $\text{supp}(p)$ does not contain $F_t$. This means that there is a linear inequality $\langle g, t \rangle \geq c$ that is valid on $P$ and such that (i) $\langle g, f_i \rangle = c$ for all $i \in F_t$; and (ii) $\langle g, f_i \rangle > c$ for some $i \in \text{supp}(p)$. 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 $\text{supp}(p_\ast) \subseteq F_t$ and finishes the proof of $\text{supp}(p_\ast) = F_t$.

We have shown the two properties, and it remains to argue that the EMLE is unique. But this follows from the fact that $\text{supp}(p_\ast)$ is equal to $F_t$, and $l_F$ is strictly convex (up to identifiability), such that there is a unique distribution on $E_{\Delta, F_t}$ that maximizes the likelihood.

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

I. Csiszár and F. Matúš. Generalized maximum likelihood estimates for exponential families. *Probability Theory and Related Fields*, 141:213–246, 2008.

I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory. now Publishers, first edition, 2004.

M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 3642042945, 9783642042942.

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

A. Dobra, E. Erosheva, and S. 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.

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. Fienberg. *The analysis of cross-classified categorical data*. Springer-Verlag, New York, second edition, 2007.
S. 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. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40:996–1023, 2012.

E. Gawrilow and M. Joswig. polymake: a framework for analyzing convex polytopes. In Gil Kalai and Günter M. Ziegler, editors, *Polytopes — Combinatorics and Computation*, pages 43–74. Birkhäuser, 2000.

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

S. Haberman. *The Analysis of Frequency Data*. Univ. Chicago Press, Chicago, IL, 1974.

V. Karwa and A. Slavkovic. Inference using noisy degrees. *Annals of Statistics*, 44:87–112, 2016.

S. 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.

Q. Liu and A. Ihler. Distributed parameter estimation via pseudo-likelihood. *International Conference on Machine Learning ICML*, 2012.

H. Massam and N. Wang. A local approach to estimation in discrete loglinear models. *Preprint*, 2015. arXiv:1504.05434.

H. Massam and N. Wang. Local conditional and marginal approach to parameter estimation in discrete graphical models. *Journal of Multivariate Analysis*, 164:1–21, 2017.

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. Wainwright, and J. Lafferty. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

M. Schmidt. minFunc: unconstrained differentiable multivariate optimization in Matlab. *http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html*, 2005.

M. Vlach. Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Discrete Applied Mathematics*, 13(1):61–78, 1986.

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

E-MAIL: wangnanw@yorku.ca E-MAIL: jrauh@mis.mpg.de

E-MAIL: massamh@yorku.ca