The Annals of Statistics
2019, Vol. 47, No. 3, 1203–1233
https://doi.org/10.1214/18-AOS1710
© Institute of Mathematical Statistics, 2019

APPROXIMATING FACES OF MARGINAL POLYTOPES IN DISCRETE HIERARCHICAL MODELS

BY NANWEI WANG[*], JOHANNES RAUH[*,†] AND HÉLÈNE MASSAM[*,1]

York University[*] and Max Planck Institute for Mathematics in the Sciences[†]

The existence of the maximum likelihood estimate in a hierarchical log-linear model 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 $\mathbf{F}_t$ 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 $\mathbf{F}_t$ exactly. Massam and Wang (2015) found an outer approximation to $\mathbf{F}_t$ using a collection of submodels 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 $\mathbf{F}_t$.

Knowing $\mathbf{F}_t$ exactly indicates which cell probabilities have maximum likelihood estimates equal to 0. When $\mathbf{F}_t$ cannot be obtained exactly, we can use, first, the outer approximation $\mathbf{F}_2$ to reduce the dimension of the problem and then the inner approximation $\mathbf{F}_1$ to obtain correct estimates of cell probabilities corresponding to elements of $\mathbf{F}_1$ and improve the estimates of the remaining probabilities corresponding to elements in $\mathbf{F}_2 \setminus \mathbf{F}_1$. Using both real-world and simulated data, we illustrate our results, and show that our methodology scales to high dimensions.

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 nonexistence 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 nonexistence of the MLE,

Received March 2016; revised April 2018.

[1] Supported by an NSERC Discovery Grant.

MSC2010 subject classifications. 62H12, 62F10, 62F99.

Key words and phrases. Existence of the maximum likelihood estimate, marginal polytope, faces, facial sets, extended maximum likelihood estimate.

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 (log-linear 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 $\mathcal{I}_+ = \{i_1, \dots, i_N\}$ be the outcome of N independent repetitions of the experiment within either a multinomial or Poisson setting. The data $\mathcal{I}_+$ is summarized by the vector t of sufficient statistics, which is of the form $t = \sum_{j=1}^N f_{i_j}$ for some vectors $f_i, i \in I$, determined by the given hierarchical log-linear model. Under these assumptions, the distribution of the data belongs to a natural exponential family with density

$$f(i_1, \dots, i_N; \theta) = \exp\{\langle \theta, t\rangle - Nk(\theta)\} \tag{1}$$

with respect to the counting measure, where θ is a log-linear parameter. To each exponential family is associated a polytope $\mathbf{P}$, called the marginal polytope, which is the convex hull of the vectors $f_i, i \in I$. Furthermore, $\mathbf{P}$ 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 $\mathbf{P}$. If the MLE does not exist, then $\frac{t}{N}$ belongs to the relative interior of a face denoted $\mathbf{F}_t$. It is the smallest face of $\mathbf{P}$ containing $\frac{t}{N}$, and it is proper (i.e., $\mathbf{F}_t \neq \mathbf{P}$). Thus, determining whether, for a given data set, the MLE of the parameter of a discrete hierarchical log-linear model exists is equivalent to determining whether $\frac{t}{N}$ belongs to a proper face of $\mathbf{P}$. The parameter may be the log-linear parameter θ, or the cell probabilities $p = (p(i), i \in I)$ obeying the constraints of the model and in 1–1 correspondence with θ. The MLE can thus be thought of in terms of θ, or in terms of p.

If the MLE does not exist, it is still possible to compute the extended MLE (EMLE) [Barndorff-Nielsen (1978), Csiszár and Matúš (2008), Lauritzen (1996)], which is a probability distribution that maximizes the likelihood over the closure of the hierarchical model (i.e., the EMLE can be approximated arbitrarily well by distributions from the hierarchical model). The support of the EMLE is given by $F_t = \{i \in I : f_i \in \mathbf{F}_t\}$, called the facial set of $\mathbf{F}_t$. 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 $\dim(\mathbf{F}_t)$, generated by a measure with support F_t [Geyer (2009)]. Therefore, precise knowledge of $\mathbf{F}_t$ 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 $\mathbf{F}_t$. This is easy when the face lattice of $\mathbf{P}$ 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 (2010)]. 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 Slavković (2016) and papers cited therein. When the face lattice of $\mathbf{P}$ cannot be computed, algorithms to compute $\mathbf{F}_t$ 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 $\mathbf{F}_t$ by relating it to faces of smaller hierarchical models as follows. A hierarchical model for the discrete random vector $X = (X_v, v \in V)$ is determined by a set of interactions among its components X_v, $v \in V$, that is represented by a simplicial complex Δ. Massam and Wang (2015) consider subsets V_i, $i = 1, \dots, k$, 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 $\mathbf{F}_{t_i}$ containing the corresponding sufficient statistic t_i, $i = 1, \dots, k$. 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 $\mathbf{F}_2$ of $\mathbf{P}$ containing $\mathbf{F}_t$. Thus, if $\mathbf{F}_2$ is a proper face of $\mathbf{P}$, then $\mathbf{F}_t$ is also a proper face and, therefore, the MLE does not exist. While Massam and Wang (2015) work with graphical models, we show that their results also hold for hierarchical log-linear models.

We call $\mathbf{F}_2$ an *outer approximation* to $\mathbf{F}_t$. 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 $\mathbf{F}_2$ does not necessarily have the same dimension as $\mathbf{F}_t$.

The purpose of this paper is to add to this outer approximation $\mathbf{F}_2$ an inner approximation $\mathbf{F}_1$ that is a subset of $\mathbf{F}_t$. While $\mathbf{F}_2$ is derived from looking at simplicial subcomplexes of Δ, 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 subsimplices corresponding to its components with a small number of vertices in V. Thus, both $\mathbf{F}_1$ and $\mathbf{F}_2$ can be obtained by computing facial sets on smaller hierarchical models involving fewer nodes.

The inner and outer approximations satisfy $\mathbf{F}_1 \subseteq \mathbf{F}_t \subseteq \mathbf{F}_2$. Clearly, we want $\mathbf{F}_1$ as large as possible and $\mathbf{F}_2$ as small as possible to have as much information about $\mathbf{F}_t$ as possible. In our simulations, we observe that $\mathbf{F}_t = \mathbf{F}_2$ most of the time and that $\mathbf{F}_t = \mathbf{F}_1$ quite often. The approximations $\mathbf{F}_1$ and $\mathbf{F}_2$ allow to bound the dimension of $\mathbf{F}_t$. This can be taken into account whenever the dimension of $\mathbf{F}_t$ 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 θ, some of its components θ_j may still be finite and well defined in this situation. In Section 4, we introduce a log-linear parametrization μ, different from θ, 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 $\mathbf{F}_t$ is unknown, the parametrization μ can be adjusted to incorporate knowledge that is available in the form of inner and outer approximations $\mathbf{F}_1$ and $\mathbf{F}_2$.

We extend the work of Fienberg and Rinaldo (2012) and that of Geyer (2009) in several directions: first, we construct approximations to $\mathbf{F}_t$ in high dimensions when a direct computation of $\mathbf{F}_t$ is not feasible. Second, we explicitly identify all parameter combinations that remain finite and meaningful when the MLE does not exist and $\mathbf{F}_t$ is known, and we also discuss what can be said when only approximations to $\mathbf{F}_t$ 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 $\mathbf{F}_1$ and $\mathbf{F}_2$. In Section 4, we show how to use $\mathbf{F}_1$ and $\mathbf{F}_2$ 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, Erosheva and Fienberg (2004) and Dobra and Lenkoski (2011), illustrates how the outer approximation $\mathbf{F}_2$ 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 × 10 grid, and the real-world data set of voting records in the US Senate.

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, 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, Kahle and Ay (2011). Let $X = (X_v, v \in V)$ be a discrete random vector with components indexed by a finite set V. 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 π_D.

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

$$\log p(i) = \sum_{D \in \Delta} \theta_D(i_D), \tag{2}$$

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 θ, we also write $p_\theta(i)$ instead of $p(i)$. The set of all such distributions $\mathcal{E}_\Delta := \{p_\theta\}$ is called the *hierarchical model* of Δ.

Equation (2) is 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

$$\log p_\theta(i) = \sum_{j \in J} \theta_j a_{j,i} - k(\theta), \tag{3}$$

where $A_\Delta = (a_{j,i})_{j \in J, i \in I}$ is a fixed real matrix (depending on Δ) and where $k(\theta) = \log \sum_i \exp(\sum_i \theta_j a_{j,i})$ ensures that $\sum_{i \in I} p_\theta(i) = 1$. 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 Supplementary Material Appendix A [Wang, Johannes and Massam (2018)].

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 $i, j \in D$, $i \neq j$, 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 (1980), Rauh, Kahle and Ay (2011). Generalizing (3), let I and J be finite sets and let $A \in \mathbf{R}^{J \times I}$ be a real matrix. Denote the columns of A by f_i, $i \in I$. The discrete exponential family corresponding to A, denoted by $\mathcal{E}_A$, consists of all probability distributions on I of the form

$$p_\theta(i) = \exp\{\langle\theta, f_i\rangle - k(\theta)\} = \exp\{(A^t\theta)_i - k(\theta)\}, \qquad \theta \in \mathbf{R}^J, \tag{4}$$

where, as above, $k(\theta) = \log\sum_i \exp(\sum_j \theta_j a_{j,i})$. It is convenient to write $\tilde{A}$ for the $(1 + |J|) \times I$ matrix with columns equal to $\binom{1}{f_i}$, $i \in I$, and to set $\theta_0 := -k(\theta)$ and $\tilde{\theta} = (\theta_0, \theta)$ (as a column vector). Then (4) rewrites to

$$p_\theta(i) = \exp(\tilde{A}^t\tilde{\theta}), \qquad \theta \in \mathbf{R}^J. \tag{5}$$

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 $\mathbf{P}_A$. In the case of a hierarchical model, $\mathbf{P}_\Delta := \mathbf{P}_{A_\Delta}$ 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 $h' = \operatorname{rank}(A_0) = \operatorname{rank}(\tilde{A}_0) - 1$, 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 μ_L are identifiable parameters on $\mathcal{E}_A$: in fact, their number is equal to h', and they are independent by construction.

The definition of $\mu_i(\theta)$ can be extended to all $i \in I$. However, only the μ_i with $i \in L$ are free parameters, while the μ_i with $i \in I \setminus L$ are linear functions of μ_L. The μ_i can be interpreted as log-likelihood ratios:

$$\mu_i(\theta) = \log\frac{p_\theta(i)}{p_\theta(0)}, \qquad \mu_0(\theta) = 0.$$

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

$$\tilde{A}n = \begin{pmatrix} N \\ t \end{pmatrix} \quad \text{and} \quad An = t, \tag{6}$$

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

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

The log-likelihood function for the log-linear parameters θ of $\mathcal{E}_A$ is therefore

$$l(\theta) = \sum_{j \in J} \theta_j t_j - Nk(\theta). \tag{7}$$

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

$$l(\mu) = \sum_{i \in I} n(i) \log p(i) = \sum_{i \in I} n(i) \mu_i - N \log\Bigl(\sum_{i \in I} \exp \mu_i\Bigr). \tag{8}$$

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

2.2. *The convex support and its facial sets.* We next recall some facts about facial sets. We refer to Ziegler (1995) for a general introduction to polytopes and their face lattices.

The convex support polytope $\mathbf{P}_A$ 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 $\mathbf{F}$. Thus, we describe a face $\mathbf{F}$ by identifying the corresponding *facial set* $F = \{i \in I : f_i \in \mathbf{F}\}$. 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 $\mathbf{P}_A = \mathbf{P}_\Delta$ is a marginal polytope, we abbreviate $F_{A_\Delta}(S)$ by $F_\Delta(S)$.

As mentioned in the Introduction, to derive the inner approximation $\mathbf{F}_1$ to $\mathbf{F}_t$ and its outer approximation $\mathbf{F}_2$, we need to consider submodels of a given model. When one exponential family $\mathcal{E}_{A'}$ is a subset of another family $\mathcal{E}_A$, then the convex support polytope $\mathbf{P}_{A'}$ is a linear projection of $\mathbf{P}_A$, 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 (1995). For hierarchical models, these facts are summarized in the following result.

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

REMARK 2.2. It is convenient to embed $\mathbf{P}_A$ in a vector space with one additional dimension using a map $\mathbf{R}^h \to \mathbf{R}^{h+1}$, $t \mapsto \tilde{t} := (1, t)$. 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 $\mathbf{F}$ is given, its facial set F can be obtained by checking whether $f_i \in \mathbf{F}$ for each $i \in I$. In the other direction, when a facial set F is given, it is much more difficult to compute a defining inequality of the corresponding face $\mathbf{F}$. However, it is straightforward to compute the linear equations defining **F**: the set of such equations $0 = \langle g, x\rangle - c = \langle \tilde{g}, \tilde{x}\rangle$ corresponds to the set of vectors $\tilde{g} \in \ker \tilde{A}^t_F$, 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 $\mathcal{E}_A$ and cell counts $n = (n(i) : i \in I)$ given as above, a parameter value θ^* is an MLE if it is a global maximum of $l(\theta)$. A MLE need not exist, since the domain of the parameters θ 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 θ^* is an MLE if and only if p_{θ^*} maximizes $\tilde{l}$ subject to the constraint that $p \in \mathcal{E}_A$. When $\tilde{l}$ has no maximum on $\mathcal{E}_A$, we can pass to the topological closure $\overline{\mathcal{E}_A}$. It can be characterized in terms of the convex support polytope $\mathbf{P}_A$ and its facial sets as follows.

THEOREM 2.3 (Barndorff-Nielsen (1978)). *The topological closure of $\mathcal{E}_A$ is $\overline{\mathcal{E}_A} = \bigcup_F \mathcal{E}_{F,A}$, where F runs over all facial sets of $\mathbf{P}_A$ and where $\mathcal{E}_{F,A}$ consists of all probability distributions of the form $p_{F,\theta}$, with*

$$p_{F,\theta}(i) = \begin{cases} \exp(\langle \theta, f_i\rangle - k_F(\theta)) & \textit{if } i \in F, \\ 0 & \textit{otherwise}, \end{cases} \tag{9}$$

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

Thus, $\overline{\mathcal{E}_A}$ is a finite union of sets $\mathcal{E}_{F,A}$ that are exponential families themselves with a very similar parametrization, using the same number of parameters. The design matrix of $\mathcal{E}_{F,A}$ 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 $\tilde{A}_F$ never has full rank.

Although the parameters θ on $\mathcal{E}_A$ and the parameters θ on $\mathcal{E}_{F,A}$ look similar, they behave differently in the following sense: if $\theta^{(s)}$ is a sequence of parameters with $p_{\theta^{(s)}} \to p_{F,\theta}$ for some θ, then, in general, $\lim_{s\to\infty} \theta^{(s)} \neq \theta$.

THEOREM 2.4 (Barndorff-Nielsen (1978)). *For any vector of observed counts n, there is a unique maximum p^* of $\tilde{l}$ in $\overline{\mathcal{E}_A}$. This maximum p^* satisfies*: (1) $Ap^* = \frac{t}{N}$, *where $t = An$*, (2) $\operatorname{supp}(p^*) = F_t$, (3) $p^* \in \mathcal{E}_{F_t,A}$.

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 can be computed by computing the MLE on $\mathcal{E}_{F_t,A}$. If the MLE θ^* exists, then $p^* = p_{\theta^*}$.

2.4. *Decomposable models.* Computing $\mathbf{F}_t$ or finding an approximation is easier when the simplicial complex Δ of the model is decomposable. We need the following definitions.

Let $V' \subseteq V$. 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 Δ.

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

A hierarchical model is *decomposable* if its generating set is a union $\Delta = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_r$ of induced subcomplexes $\Delta_i = \Delta|_{V_i}$ in such a way that:

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

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

Lemma 2.5 below states that, if Δ is reducible, then any facial set for Δ 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 Δ be reducible into two components $\Delta|_{V_1}$ and $\Delta|_{V_2}$:*

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

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

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

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

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

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

$$F_\Delta(T) = \pi_1^{-1}\bigl(\pi_1(T)\bigr) \cap \pi_2^{-1}\bigl(\pi_2(T)\bigr) \cap \cdots \cap \pi_r^{-1}\bigl(\pi_r(T)\bigr) \tag{10}$$

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 Δ and marginal polytope $\mathbf{P}_\Delta$. 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 Δ and to complete it. To be precise, we augment Δ by adding all subsets of S. 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 is an 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 Δ. Its facial sets can be computed from (10).

In general, the approximation obtained from a single separator (or, in general, a single supercomplex) is not good; that is, $F_t = F_\Delta(I_+)$ 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

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

that satisfy

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

where all inclusions except the last one are due to the definition of $F_{\Delta_S}(T)$ or $F_{\Delta_{S'}}(T)$ as the smallest facial sets containing T in Δ_S or $\Delta_{S'}$. The last inclusion is a consequence of Lemma 2.1 since both Δ_S and $\Delta_{S'}$ contain Δ. 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 Δ_S and $\Delta_{S'}$. The same iteration can be done replacing Δ_S and $\Delta_{S'}$ by Δ_{V_1,V_2} and $\Delta_{V'_1,V'_2}$. 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 Δ_{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. It has two problems: 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 neighbors from all other nodes. In this case, V_1 consists of i and its neighbors, 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_{Δ_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 (e.g., $\min\{|V_1 \setminus S|, |V_2 \setminus S|\} \geq 3$).

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 subcomplex $\Delta' \subseteq \Delta$ provides an outer approximation of $F_\Delta(S)$. Removing sets from Δ 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 subcomplexes.

When comparing Δ with an induced subcomplex $\Delta|_{V'}$ for some $V' \subseteq V$, 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 $i' \in I_{V'}$,

$$f_i = f'_{i'} \quad \Leftrightarrow \quad i \in \pi_{V'}^{-1}(i'). \tag{11}$$

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 any $K \subseteq I$, let $F'_{\Delta|_{V'}}(K)$ be the facial set when $\Delta|_{V'}$ is considered as a simplicial complex on V', and let $F_{\Delta|_{V'}}(K)$ be the facial set when $\Delta|_{V'}$ is considered as a simplicial complex on V. Then*

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

PROOF. For any $K \subseteq I$, the set $\mathcal{A} = \{f_i, i \in K\}$ is identical to $\mathcal{B} = \{f'_{i'}, i' \in \pi_{V'}(K)\}$. Therefore, the smallest faces of the marginal polytopes for $\Delta_{V'}$ on V or V' containing $\mathcal{A}$ and $\mathcal{B}$, respectively, are the same.

By definition of $F'_{\Delta_{V'}}(\pi_{V'}(K))$, the smallest face containing $\mathcal{B}$ is defined by $\{f'_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$. By definition of $F_{\Delta_{V'}}(K)$, the smallest face containing $\mathcal{A}$ is $\{f_i, i \in F_{\Delta_{V'}}(K)\}$. Also, $\{f_i, i \in \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))\} = \{f'_{i'}, i' \in F'_{\Delta_{V'}}(\pi_{V'}(K))\}$ by (11). Thus, $F_{\Delta_{V'}}(K) = \pi_{V'}^{-1}(F'_{\Delta_{V'}}(\pi_{V'}(K)))$. □

In general, $F_{\Delta|_{V'}}(I_+)$ is not a good approximation of $F_\Delta(I_+)$. It can be improved by considering several subsets of $V_1, \dots, V_r \subseteq V$. Then $F_\Delta(I_+) \subseteq F_{\Delta|_{V_i}}(I_+)$, $i = 1, \dots, r$, and so $F_\Delta(I_+) \subseteq \bigcap_{i=1}^r F_{\Delta|_{V_i}}(I_+) =: F_{V_1,\dots,V_r;\Delta}(I_+)$. 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 Δ, 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 $F_{V_1,\dots,V_r;\Delta}(I_+)$. 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 Δ. The parameter k should be chosen as large as possible such that computing $F_{V_1,\dots,V_r;\Delta}(I_+)$ is still feasible. Note that computing $F_{\Delta|_D}(I_+)$ for $D \in \Delta$ is trivial, since $\mathbf{P}_{\Delta|_D}$ 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 Δ. For example, for the grid graphs, we suggest to use 3×3 subgrids. 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 3×3 subgrid that contains v and all neighbors of v. We will compare two different strategies:

3. for a grid, use all 3×3 subgrids;
4. cover a grid by 3×3 subgrids.

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

In general, it is not enough to look at induced subcomplexes, unless Δ 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 $F_1 \subseteq F_t \subseteq F_2$. In the lucky case that $F_1 =$

F_2, we know that $F_t = F_1 = F_2$. In general, the cardinality of $F_2 \setminus F_1$ indicates the quality of our approximations.

F_1, F_2 and F_t can also be compared by the ranks of the matrices $\tilde{A}_{F_1}$, $\tilde{A}_{F_2}$ and $\tilde{A}_{F_t}$ obtained from $\tilde{A}$ by keeping only the columns indexed by F_1, F_2 and F_t, respectively. Clearly, $\operatorname{rank} \tilde{A}_{F_1} \le \operatorname{rank} \tilde{A}_{F_t} \le \operatorname{rank} \tilde{A}_{F_2}$. Note that $\operatorname{rank} \tilde{A}_{F_2} - 1$ equals the dimension of the corresponding face $\mathbf{F}_2$ of $\mathbf{P}$, and $\operatorname{rank} \tilde{A}_{F_t} - 1$ equals the dimension of $\mathbf{F}_t$. Although F_1 does not necessarily correspond to a face of $\mathbf{P}$, we can bound the codimension of $\mathbf{F}_t$ in $\mathbf{F}_2$ by

$$\dim \mathbf{F}_2 - \dim \mathbf{F}_t \le \operatorname{rank} \tilde{A}_{F_2} - \operatorname{rank} \tilde{A}_{F_1}.$$

In particular, if $\operatorname{rank} \tilde{A}_{F_2} = \operatorname{rank} \tilde{A}_{F_1}$, 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.* When the MLE exists, it can be computed by numerically maximizing the log-likelihood function $l(\theta)$ given in (7). As mentioned before, $l(\theta)$ is concave (or even strictly concave, if the parameters θ 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 $k(\theta)$; 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. There are also dedicated algorithms, such as *iterative proportional fitting* (IPF), which is of Gauss–Seidel type [Csiszár and Shields (2004)].

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

$$l_{F_t}(\theta) = \log\Biggl(\prod_{i \in I_+} p_{F_t,\theta}(i)^{n(i)}\Biggr) = \sum_{j \in J} \theta_j t_j - N k_{F_t}(\theta). \tag{12}$$

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_t is smaller than I. However, as stated above, the parametrization $\theta \mapsto p_{F_t,\theta}$ is never identifiable. Of course, this problem is easy to solve by selecting a set of independent parameters among the θ_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 θ_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 $\mathcal{E}_{F_t,A}$, 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, the maximum is not at a finite value θ^*, 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 toward the right direction. Such an algorithm generates a sequence of parameter values $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \dots$ with increasing log-likelihood values $l(\theta^{(1)}) \le l(\theta^{(2)}) \le \dots$. 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 $\sup_\theta l(\theta) = \max_{p\in\overline{\mathcal{E}_A}} \tilde{l}(p)$. The algorithm will stop when the difference $l(\theta^{(s+1)}) - l(\theta^{(s)})$ becomes negligeably 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 θ_j will tend to ±∞, 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 Supplementary Material Appendix B [Wang, Johannes and Massam (2018)] 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, that is, 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 behavior of the remaining parameters, but, as will be illustrated in Section 5.2, the MLE obtained with the model on $\mathbf{F}_2$ 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 parameter estimation problems mentioned above.

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

1. The parametrization is not identifiable, that is, there are parameters θ, θ' with $p_{F_t,\theta} = p_{F_t,\theta'}$.

2. While the parametrization $\theta \mapsto p_{F_t,\theta}$ of $\mathcal{E}_{F_t,A}$ looks similar to the parametrization $\theta \mapsto p_\theta$ of $\mathcal{E}_A$, the values of the parameters in both parametrizations are not related to each other.

3. When $p_{\theta^{(s)}} \to p_{F_t,\theta}$ as $s \to \infty$ for some parameter values $\theta^{(s)}$, θ, 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 μ_L (introduced in Section 2.1) solve 1 and 2 and improve 3. Afterward, we discuss what can be done if F_t is not known. We briefly discuss the general solution toward 3 in Supplementary Material Appendix C [Wang, Johannes and Massam (2018)]. In any case, the choice of the parameters will depend on the 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 μ_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 μ_i are not independent, and we need to choose an independent subset L. We do this in two steps:

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

2. Then extend L_t to a maximal subset $L \subseteq I$ such that the parameters μ_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 μ_i, $i \in L_t$, of the parameters μ_L gives an identifiable parametrization of $\mathcal{E}_{F_t,A}$.

2. Let μ_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 $\mu_i^{(s)}$ are such that

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

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

The last property ensures a consistency of the parameters μ_i on $\mathcal{E}_A$ and on $\mathcal{E}_{F_t,A}$. 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 μ_i, but the original parameters θ_i that have an interpretation. But when we understand the parameters μ_i, we can also tell which of the paramters θ_i or which combinations of the parameters θ_i have finite well-defined values and can be computed, and which parameters diverge.

LEMMA 4.1. *Let $\theta^{(s)}$, $s \in \mathbb{N}$, be parameter values such that $p_{\theta^{(s)}} \to p^*$ as $s \to \infty$. For any $i \in L_t$, the linear combination $\mu_i^{(s)} = \langle \theta^{(s)}, f_i - f_0 \rangle$ has a well-defined finite limit as $s \to \infty$. Any linear combination of the $\theta_i^{(s)}$ with a well-defined finite limit* (*i.e.*, *a limit that is independent of the choice of the sequence* $\theta^{(s)}$) *is a linear-combination of the $\mu_i^{(s)}$ 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 θ is also a linear combination of the μ, 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 μ_j with $j \notin L_t$, then there exist sequences $\mu^{(s)}$, $\mu'^{(s)}$ of parameters with

$$\lim_{s\to\infty} p_{\mu^{(s)}} = \lim_{s\to\infty} p_{\mu'^{(s)}} \quad \text{and} \quad \lim_{s\to\infty} \sum_i a_i \mu_i^{(s)} \neq \lim_{s\to\infty} \sum_i a_i \mu_i'^{(s)}.$$

So suppose that $\mu^{(s)}$ is a sequence of parameters such that $\lim_{s\to\infty} p_{\mu^{(s)}}$ exists and such that $\lim_{s\to\infty} \sum_i a_i \mu_i^{(s)}$ is finite. Define

$$\mu_i'^{(s)} = \begin{cases} \mu_j^{(s)} + 1 & \text{if } i = j, \\ \mu_i^{(s)} & \text{otherwise.} \end{cases}$$

Then an easy computation shows that $\lim_{s\to\infty} p_{\mu'^{(s)}} = \lim_{s\to\infty} p_{\mu^{(s)}}$ and $\lim_{s\to\infty} \sum_i a_i \mu_i'^{(s)} = \lim_{s\to\infty} \sum_i a_i \mu_i^{(s)} + a_j$. □

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 μ_i:

1. Choose a maximal subset L_1 of F_1 such that the parameters μ_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 μ_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 μ_i, $i \in L$ are independent by adding elements $i \in I \setminus F_2$.

The following properties 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 $\mu_i^{(s)} = \langle \theta^{(s)}, f_i - f_0 \rangle$:*

1. *For any $i \in L_1$, the linear combination $\mu_i^{(s)} = \langle \theta, f_i - f_0 \rangle$ 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 Section 5.1, we simulate data for the graphical model of the 4 × 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 Section 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 × 4 *grid graph.* We generated random samples of varying sizes for the graphical model of the 4 × 4 grid graph with binary variables [Figure 1(a)]. 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

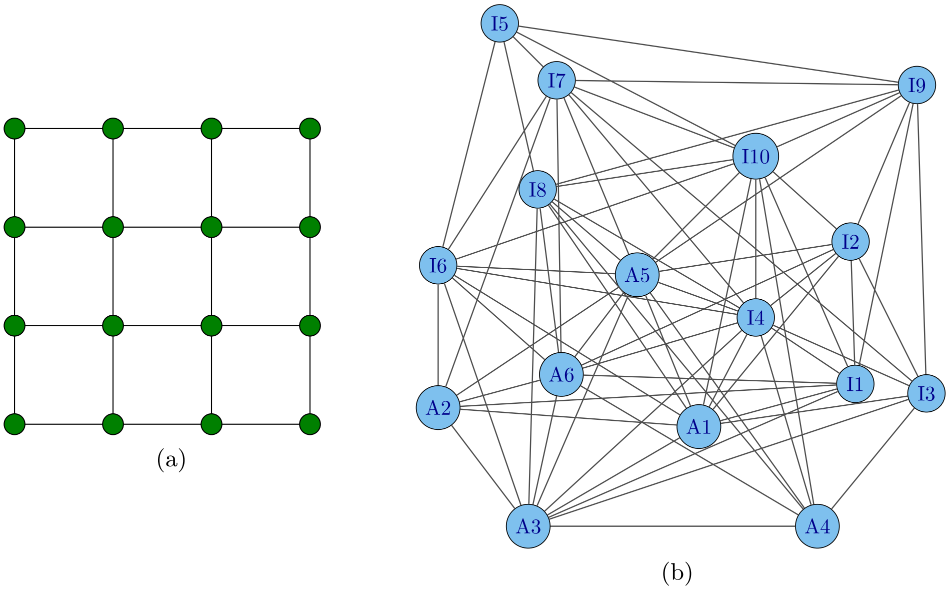


FIG. 1. (a) 4 × 4 *grid graph. Graphical model for NLTCS data set.* (b) *The label "A*n*" abbreviates ADL*n*, "I*n*" abbreviates IADL*n*.*

TABLE 1
Facial set approximation of 4×4 *grid graph*(*hierarchical log-linear model with parameters from standard normal distribution*)

Sample size	MLE does not exist	$F_1 = F_t$	$F_2 = F_t$
10	100.0%	97.7%	100.0%
50	89.5%	100.0%	100.0%
100	71.0%	100.0%	100.0%
150	52.0%	100.0%	100.0%

all possible separators, of which there are 106 (Strategy 3.1 in Section 3.1), or we iterate over the 3 horizontal, 3 vertical and 8 diagonal separators only (Strategy 3.1 in Section 3.1). We obtain the same result with either strategy. Clearly, Strategy 3.1 is much faster. To compute the outer approximation, we cover the 4×4 grid by four 3×3 grids (Strategy 3.2 in Section 3.2).

We generate samples from the hierarchical model P_θ, where the parameter vector θ 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 θ is χ^2-distributed with 40 degrees of freedom (since the number of parameters is 40). Thus, the expected squared length of θ is 40, which is large enough to move the distribution p_θ close to the boundary of the model. Indeed, we observed that when the MLE does not exist, the squared length of the numerical estimate of the MLE vector is of the order of magnitude of 40 (see also the example in Section 5.2). In all samples that we generated, $F_t = F_2$, 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 Supplementary Material Appendix D [Wang, Johannes and Massam (2018)].

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 1(b). 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.

To compare the estimates obtained with or without worrying about the existence of the MLE and with or without an approximation to F_t, we maximize the log likelihood given in terms of μ, rather than θ, as in (8). First, we ignore the fact that the MLE might not exist and numerically optimize the likelihood directly: we call this estimate $\hat{\mu}^{\mathrm{MLE}}$. Second, we find F_t and compute the EMLE with parameters denoted $\hat{\mu}^{\mathrm{EMLE}}$. Third, we obtain an inner and outer approximation to F_t and consider the resulting information on likelihood maximization. We call the resulting estimate $\hat{\mu}^{F_2' \setminus F_1'}$. All estimates are computed using the MATLAB function minFunc [Schmidt (2005)].

We first compute the inner approximation F_1 that makes use of all the separators in the graph (Strategy 3.1 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 3.2 in Section 3.2). We obtain $F_1 = F_2$, and thus deduce that $F_t = F_1 = F_2$. We find $|F_t| = 49{,}536$. Therefore, $|F_t^c| = 2^{16} - 49{,}536 = 16{,}000$ 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 log likelihood function l_{F_t} as in (12). Since $\operatorname{rank}(\tilde{A}_{F_t}) = 303$, the dimension of $\mathbf{F}_t$ 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 log likelihood, 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 $\dim \mathbf{F}_1' = \operatorname{rank} \tilde{A}_{F_1'} - 1 = 300$. 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 $\dim \mathbf{F}_2' = \operatorname{rank} \tilde{A}_{F_2'} - 1 = 310$. 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

$$l_{F_2'}(\mu) = \sum_{i \in I_+} \mu_i n(i) - N \sum_{i \in F_2'} \exp(\mu_i). \tag{13}$$

We know that μ_i is estimable for $i \in F_1'$, that μ_i goes to negative infinity for $i \in F_2'^c$, and we cannot say anything for μ_i with $i \in F_2' \setminus F_1'$.

As explained in Section 4.2, the components of μ 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 μ_i, $i \in F_2'$, 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

$$l_{F_2'}(\mu_{L_2}) = \sum_{i\in I_+} \langle b_i, \mu_{L_2}\rangle n(i) - N \sum_{i\in F_2'} \exp\langle b_i, \mu_{L_2}\rangle. \tag{14}$$

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 $\hat{\mu}_i^{F_2'\setminus F_1'} := \langle b_i, \hat{\mu}_{L_2}\rangle$, $i \in F_2'$. In total, there are $|L_2| = \operatorname{rank}(\tilde{A}_{F_2'}) - 1 = 310$ independent parameters in $l_{F_2'}$. Among them, there are $|L_1| = \operatorname{rank}(\tilde{A}_{F_1'}) - 1 = 300$ estimable parameters μ_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 $\hat{\mu}_i^{\mathrm{MLE}}$, $\hat{\mu}_i^{\mathrm{EMLE}}$ and $\hat{\mu}_i^{F_2'\setminus F_1'}$ for 19 arbitrarily chosen parameters among the 310 possible ones. The naive estimator $\log \frac{n_i}{n_0}$ is also listed. The first column 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 μ_i with a finite estimate are those for $i \in F_t$. This is illustrated in the $\hat{\mu}^{\mathrm{EMLE}}$ column, with finite values for $\hat{\mu}_i^{\mathrm{EMLE}}$, $i \in F_t$ (green and pink rows), and infinite values for $\hat{\mu}_i^{\mathrm{EMLE}}$, $i \in I \setminus F_t$ (yellow and blue rows). The last column

TABLE 2

Parameter estimates from 3 *methods compared with the relative frequency in the NLTCS data. Here, each* $i = (i_1, \dots, i_{16}) \in I = \{0, 1\}^{16}$ *is represented by the natural number* $\sum_{j=1}^{16} i_j 2^{j-1} \in \{0, \dots, 2^{16} - 1\}$

		Naive estimate	Maximum	Likelihood	Estimates
	Parameter	$\log n_i/n_0$	$\hat{\mu}_i^{\mathrm{MLE}}$	$\hat{\mu}_i^{\mathrm{EMLE}}$	$\hat{\mu}_i^{F_2'\setminus F_1'}$
$i \in F_1'$	μ_{512}	-1.2472	-1.2482	-1.2482	-1.2482
	μ_{65536}	-1.7644	-1.7976	-1.7975	-1.7975
	μ_{16}	-2.3958	-2.3844	-2.3846	-2.3846
	μ_{528}	-2.5429	-2.6504	-2.6504	-2.6504
	μ_{2048}	-2.8813	-2.7246	-2.7243	-2.7243
$i \in F_t \setminus F_1'$	μ_{32960}	$-\infty$	-13.8205	-13.8207	-13.8205
	μ_{34881}	$-\infty$	-14.3693	-14.3693	-14.3692
$i \in F_2' \setminus F_t$	μ_{36864}	$-\infty$	-30.8729	$-\infty$	-34.9805
	μ_{36880}	$-\infty$	-39.6536	$-\infty$	-45.2229
	μ_{388}	$-\infty$	-28.9090	$-\infty$	-29.4525
	μ_{32769}	$-\infty$	-32.3799	$-\infty$	-36.9537
	μ_{385}	$-\infty$	-37.1365	$-\infty$	-35.9399
	μ_{449}	$-\infty$	-38.9673	$-\infty$	-44.9405
	μ_{32785}	$-\infty$	-40.1221	$-\infty$	-45.8318
	μ_{389}	$-\infty$	-43.7297	$-\infty$	-40.0158
$i \in I \setminus F_2'$	μ_{256}	$-\infty$	-35.5482	$-\infty$	$-\infty$
	μ_{320}	$-\infty$	-42.5454	$-\infty$	$-\infty$
	μ_{257}	$-\infty$	-52.9224	$-\infty$	$-\infty$
	μ_{321}	$-\infty$	-60.2208	$-\infty$	$-\infty$

contains the estimates $\hat{\mu}_i^{F'_2\setminus F'_1}$ obtained from numerically optimizing (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 components of $\hat{\mu}^{\mathrm{EMLE}}$. This can be seen by verifying numerically that the square length of the projection on F_t of the difference between $\hat{\mu}^{\mathrm{MLE}}$ and $\hat{\mu}^{\mathrm{EMLE}}$ is greater than that between $\hat{\mu}^{F'_2\setminus F'_1}$ and $\hat{\mu}^{\mathrm{EMLE}}$. Indeed, we have

$$\|\hat{\mu}_{F_t}^{F'_2\setminus F'_1} - \hat{\mu}_{F_t}^{\mathrm{EMLE}}\|^2 \approx 6.49 < \|\hat{\mu}_{F_t}^{\mathrm{MLE}} - \hat{\mu}_{F_t}^{\mathrm{EMLE}}\|^2 \approx 8.52.$$

The components $\hat{\mu}_i^{F'_2\setminus F'_1}$ indexed by $i \in F'_2 \setminus F_t$ are finite while the corresponding components of $\hat{\mu}^{\mathrm{EMLE}}$ are infinite but they are better than those of $\hat{\mu}^{\mathrm{MLE}}$: numerically, we have

$$\sum_{i\in\mathbb{F}'_2\setminus F_t} (\hat{\mu}_i^{F'_2\setminus F'_1})^2 \approx 5184 > \sum_{i\in\mathbb{F}'_2\setminus F_t} (\hat{\mu}_i^{\mathrm{MLE}})^2 \approx 4752.$$

The estimates $\hat{\mu}_i^{F'_2\setminus F'_1}, i \in F'_2 \setminus F_t$ are better than the corresponding $\hat{\mu}_i^{\mathrm{MLE}}$ 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 μ_i^{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 the Supplementary Material Appendix E [Wang, Johannes and Massam (2018)].

6. Computing faces for large complexes. If our statistical model contains many variables and is not reducible, the problem of determining $\mathbf{F}_t$ 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 × 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 Δ contains a complete separator separating V into V_1 and V_2, we can identify a facial set F implicitly without listing it explicitly. We only need the two projections $F_{V_1} = \pi_{V_1}(F)$ and $F_{V_2} = \pi_{V_2}(F)$. Since $F = \pi_{V_1}^{-1}(F_{V_1}) \cap \pi_{V_2}^{-1}(F_{V_2})$ (by Lemma 2.5), these two projections identify F, and they allow us to do most of the operations that we would want to do with F. For example, for any $i \in I$, we can check whether $i \in F$ by checking whether $\pi_{V_1}(i) \in F_{V_1}$ and $\pi_{V_2}(i) \in F_{V_2}$, and we can check whether $F \supseteq I$ by checking whether $F_{V_1} \supseteq I_{V_1}$ and $F_{V_2} \supseteq I_{V_2}$. In particular, we can check whether the MLE exists by looking only at the two subsets V_1 and V_2.

Similar ideas apply if Δ contains a separator that is not complete. Suppose that S separates V_1 from V_2 in Δ. We want to use $F_2 := F_{\Delta|_{V_1}}(I_+) \cap F_{\Delta|_{V_2}}(I_+)$

as an outer approximation and $F_1 := F_{\Delta_S}(I_+)$ 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 $F_{1,V_1} := F_{\Delta_S|_{V_1}}(\pi_{V_1}(I_+))$ and $F_{1,V_2} := F_{\Delta_S|_{V_2}}(\pi_{V_2}(I_+))$. 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 on 100 nodes with no particular regularity pattern. In Section 6.2, we consider a grid graph and work with two families of "parallel" separators.

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, El Ghaoui and d'Aspremont (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 ℓ_1-regularized logistic regression method proposed by Ravikumar, Wainwright and Lafferty (2010) to identify the neighbors 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 large connected components and 14 independent nodes.

There are 309 sample points, and $|I_+| = 278$. We want to characterize the face $\mathbf{F}_t$ 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.

The Democratic party simplicial complex Δ_d consists of 26 variables, 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 inner and outer approximations. Figure 3(b) indicates (in green and purple) two separators that

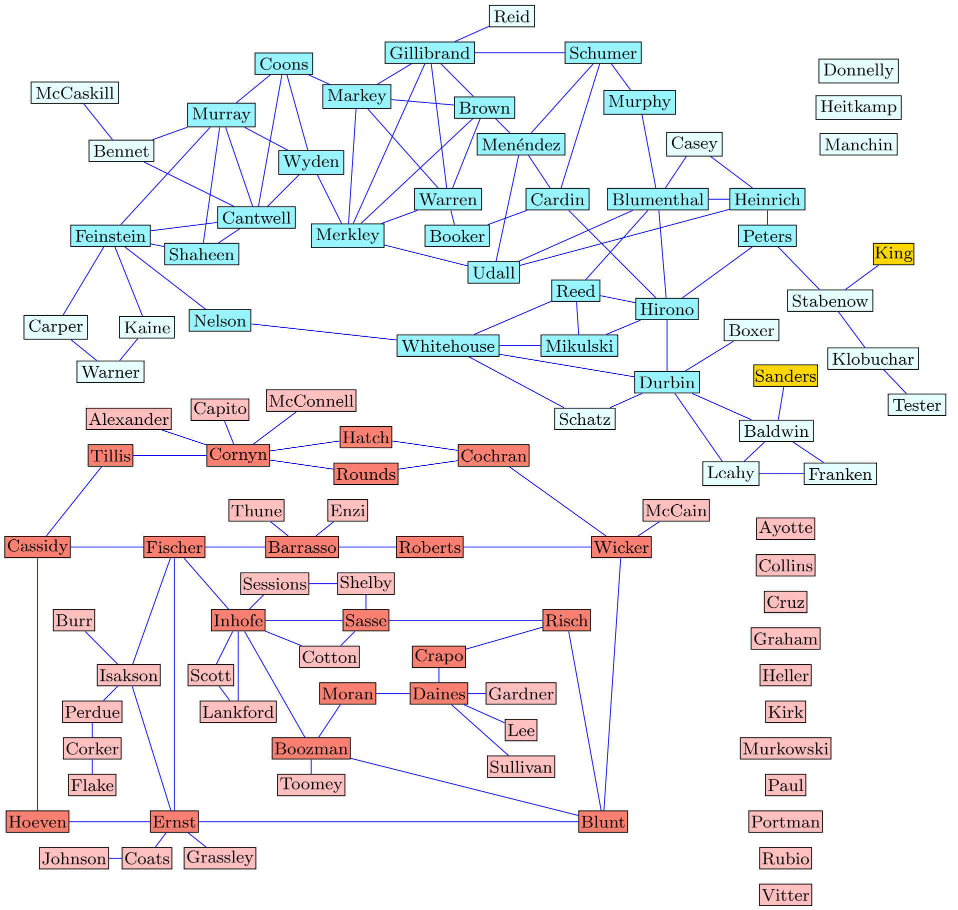
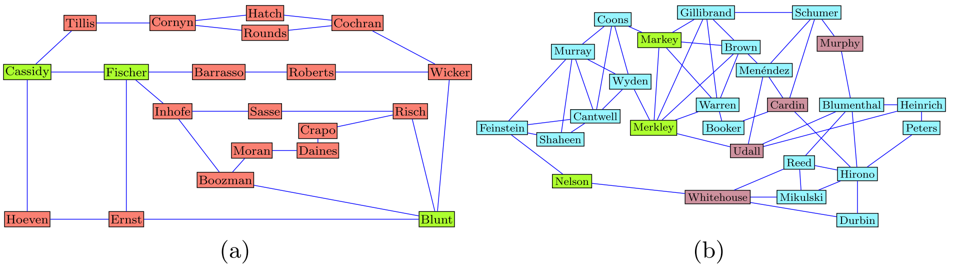


FIG. 2. *The graph for the US senate voting records data. Golden nodes are independent senators, blue nodes are Democratic and red nodes are Republican.*



(a) (b)

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

TABLE 3
Numbering of some senators

ID	Senator	ID	Senator	ID	Senator	ID	Senator
22	Nelson	37	Cardin	52	Murphy	61	Whitehouse
23	Reed	41	Markey	53	Hirono	70	Merkley
26	Schumer	47	Udall	56	Gillibrand	87	Warren

separate Δ_d into three simplicial complexes denoted, from left to right, by Δ_α, Δ_β and Δ_γ. Δ_α has 9 nodes. The corresponding vector t_α lies in the relative interior of $\mathbf{P}_{\Delta_\alpha}$. Δ_β has 13 nodes, and t_β lies on a facet $\mathbf{F}_{t_\beta}$ 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. Their numbers are given in Table 3. The inequality of $\mathbf{F}_{t_\beta}$ is

$$t_{87} - t_{56,87} \geq 0, \tag{15}$$

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." Δ_γ has 11 nodes, and t_γ lies on the facet of $\mathbf{P}_{\Delta_\gamma}$ with inequality

$$t_{23} - t_{23,53} \geq 0. \tag{16}$$

The intersection of (15) and (16) gives the outer approximation $\mathbf{F}_{2,d}$ to F_{t_d}.

To get an inner approximation, we complete each separator, that is, the green vertices are completed and the purple vertices are completed in Figure 3(b). Denote the three simplicial complexes with complete separators as $\Delta_{\tilde{\alpha}}$, $\Delta_{\tilde{\beta}}$, $\Delta_{\tilde{\gamma}}$, respectively, and let $\Delta_{\tilde{d}} = \Delta_{\tilde{\alpha}} \cup \Delta_{\tilde{\beta}} \cup \Delta_{\tilde{\gamma}}$. The smallest face $\mathbf{F}_{t_{\tilde{d}}}$ of $\mathbf{P}_{\Delta_{\tilde{d}}}$ containing $t_{\tilde{d}}$ is our inner approximation. The models of $\Delta_{\tilde{\alpha}}$, $\Delta_{\tilde{\beta}}$, $\Delta_{\tilde{\gamma}}$ and $\Delta_{\tilde{d}}$ include main effects, two-, three- and four-way interactions.

The linear programming method (applied to $\mathbf{P}_{\Delta_{\tilde{\alpha}}}$, $\mathbf{P}_{\Delta_{\tilde{\beta}}}$ and $\mathbf{P}_{\Delta_{\tilde{\gamma}}}$ separately) shows that $t_{\tilde{d}}$ belongs to the face $\mathbf{F}_{t_{\tilde{d}}}$ of $\mathbf{P}_{\Delta_{\tilde{d}}}$ with defining equations

$$\begin{aligned}
\langle g_1, t_{\tilde{d}} \rangle &= t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0, \\
\langle g_2, t_{\tilde{d}} \rangle &= t_{87} - t_{56,87} = 0, \\
\langle g_3, t_{\tilde{d}} \rangle &= t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0, \\
\langle g_4, t_{\tilde{d}} \rangle &= t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0, \\
\langle g_5, t_{\tilde{d}} \rangle &= t_{37,47,52,61} - t_{47,52,61} = 0, \\
\langle g_6, t_{\tilde{d}} \rangle &= t_{23} - t_{23,53} = 0,
\end{aligned}$$

where g_1 is contributed by $\Delta_{\tilde{\alpha}}$, $g_1, \ldots, g_5$ are contributed by $\Delta_{\tilde{\beta}}$, and g_4, g_5, g_6 are contributed by $\Delta_{\tilde{\gamma}}$. Thus, $\mathbf{F}_{1,d} := \mathbf{F}_{t_{\tilde{d}}}$ is a subset of $\mathbf{F}_{2,d}$.

A refinement of our argument shows that indeed $\mathbf{F}_{t_d} = \mathbf{F}_{2,d}$. The orthogonal complement of the subspace generated by $\mathbf{F}_{t_{\tilde{d}}}$ is

$$G = \{g' \in \mathbf{R}^{91} | g' = k_1 g_1 + k_2 g_2 + k_3 g_3 + k_4 g_4 + k_5 g_5 + k_6 g_6\}.$$

To describe $\mathbf{F}_{t_d}$, we note that each defining equation of $\mathbf{F}_{t_d}$ is of the form $\langle g, t_d \rangle = 0$, where g is orthogonal to $\mathbf{F}_{t_d}$. For any such g, let g' be its extension to a vector in $\mathbf{R}^{91}$ by adding zero components. Then $g' \perp \mathbf{F}_{t_{\tilde{d}}}$, 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 $k_1, \dots, k_5, k_6$. 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 $k_1 = k_4 = k_5 = 0$, and also the coefficient of $t_{37,52}$ must vanish, which implies $k_3 = 0$. On the other hand, the vectors g_2' and g_6' only contain interactions that are already present in Δ, and so the coefficients k_2 and k_6 are free. Thus, the equations for $\mathbf{F}_{t_d}$ are

$$(17) \qquad \langle g_2, t_{\tilde{\beta}} \rangle = t_{87} - t_{56,87} = 0, \qquad \langle g_6, t_{\tilde{\gamma}} \rangle = t_{23} - t_{23,53} = 0.$$

This is the same as the outer approximation $\mathbf{F}_{2,d}$.

The Republican simplicial complex Δ_r consists of 20 variables, and the model induced from Δ_r contains 46 parameters, which is also too large to directly compute F_{t_r}. The green nodes in Figure 3(a) separate Δ_r into two simplicial complexes denoted (from left to right) by Δ_a and Δ_b. To compute the inner approximation, we complete the green 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 $\mathbf{P}_{\Delta_{\tilde{a}}}$ and $\mathbf{P}_{\Delta_{\tilde{b}}}$, respectively. Therefore, $\mathbf{F}_1 = \mathbf{P}_{\Delta_r}$, from which we conclude that the corresponding vector t_r lies in the relative interior of $\mathbf{P}_{\Delta_r}$.

Thus, $\mathbf{F}_t$ is now determined: it is characterized by the equalities (17). What insight is there in the knowledge (i) of the nonexistence of the MLE and (ii) of $\mathbf{F}_t$? While we have given general remarks in the Introduction, let us illustrate here how knowledge about $\mathbf{F}_t$ points to some issues with the statistical analysis that would possibly be overlooked if $\mathbf{F}_t$ was not known.

First, knowing $\mathbf{F}_t$, and its defining inequalities, for one model also gives information about other models. It follows from (17) 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 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, 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 $\mathbf{F}_t$, 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 $|F_t| = \frac{9}{16} \cdot 2^{100}$ 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 combines estimates from the local conditional likelihoods derived from the distribution of each variable given its neighbors; see, for example, Liu and Ihler (2012) or Massam and Wang (2018). Thus, 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 neighbors $X_{19,53,61,78}$. 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

$$\prod_{i_a,i_b,i_c,i_d,i_e} p\bigl(X_b = i_b | X_{acde} = (i_a, i_c, i_d, i_e)\bigr)^{n(i_a,i_b,i_c,i_d,i_e)}, \tag{18}$$

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 $p(X_b = 1 \mid i_a, i_c, i_d, i_e)$ is the empirical estimate $n(i_a, i_b = 1, i_c, i_d, i_e)/n(i_a, i_c, i_d, i_e)$. 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,

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

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 maximum composite likelihood estimate is obtained by averaging the estimates obtained from various local conditional likelihoods. From the b-local conditional model, we also obtain $\hat{p}(X_b = 1 \mid i_a = 1, i_c = 1, i_d = 1, i_e = 0) = 1/2$ and $\hat{p}(X_b = 1 \mid i_a = 0, i_c = 1, i_d = 0, i_e = 1) = 4/5$, which yield the linear combinations

$$\hat{\theta}_b + \hat{\theta}_{bc} + \hat{\theta}_{ab} + \hat{\theta}_{bd} = 0, \qquad \hat{\theta}_b + \hat{\theta}_{bc} + \hat{\theta}_{be} = 1.4. \tag{19}$$

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 (18) in terms of θ in Matlab, we obtain $l_{\text{local}}(\hat{\theta}) = 3.88830675$ for $\hat{\theta} = \hat{\theta}_1 \approx (-62.3, -16.8, 35.2, 43.9, 28.5)$. If we change the starting point to $\theta_0 = (100, 100, 100, 100, 100)$, we obtain $l_{\text{local}}(\hat{\theta}_2) = 3.88830648$ and $\hat{\theta}_2 = (-162.2, -26.1, 91.2, 97.1, 72.4)$. Clearly, the values for $\hat{\theta}$ are unreliable since the MLE in the local conditional model does not exist. However, both $\hat{\theta}_1$ and $\hat{\theta}_2$ satisfy equations (19). One can, of course, obtain estimates of θ_{bc}, θ_{ab}, θ_{bd}, θ_{be} from the local conditional models centered at c, a, d and e, respectively, but these estimates are not true maximum composite likelihood estimates, and it remains to study their properties.

This example shows that our methods make it possible to obtain $\mathbf{F}_t$ for very large examples. It also illustrates how knowing $\mathbf{F}_t$ gives us precious information on the reliability of the maximum composite likelihood estimate.

6.2. *The* 5×10 *grid.* Let Δ be the simplicial complex of the 5×10 grid graph (Figure 4). 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 5×10 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 5×3 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

$$S_2 = \{11, \dots, 15\}, \qquad S_4 = \{21, \dots, 25\},$$
$$S_6 = \{31, \dots, 35\}, \qquad S_8 = \{41, \dots, 45\}$$

that separate the vertex sets $V_1 = \{1, \dots, 15\}$, $V_3 = \{11, \dots, 25\}$, $V_5 = \{21, \dots, 35\}$, $V_7 = \{31, \dots, 45\}$, $V_9 = \{41, \dots, 50\}$.

Recall that in the Senate example of Section 6.1 it was enough to work with a single family of disjoint separators to find the facial set F_t for the given vector t. However, here, the approximations F_1 and F_2 obtained with only the blue separators, say, are not tight for most data sets. Therefore, we alternate between the blue separators and the red separators

$$S_1 = \{6, \dots, 10\}, S_3 = \{16, \dots, 20\},$$
$$S_5 = \{26, \dots, 30\}, S_7 = \{36, \dots, 40\},$$

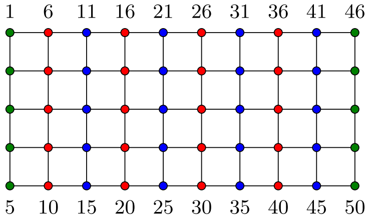


FIG. 4. 5×10 *grid graph, the red and blue nodes are the set of separators we use to compute* F_1, *they are used iteratively to get a better lower approximation.*

that separate $V_0 = \{1, \dots, 10\}$, $V_2 = \{6, \dots, 20\}$, $V_4 = \{16, \dots, 30\}$, $V_6 = \{26, \dots, 40\}$, $V_8 = \{36, \dots, 50\}$. Our inner approximation is then defined by

$$\begin{aligned} G^{(0)} &= I_+, \\ G^{(2i+1)} &= F_{\Delta_{S_2;S_4;S_6;S_8}}\big(G^{(2i)}\big), \qquad i = 0, 1, \dots \\ G^{(2i)} &= F_{\Delta_{S_1;S_3;S_5;S_7}}\big(G^{(2i-1)}\big), \qquad i = 1, 2, \dots \\ F_1 &= \bigcup_j G^{(j)}. \end{aligned}$$

As explained in Section 3.1, this recursion stabilizes and F_1 is obtained after a finite number of steps.

Since $|I| = 2^{50} \approx 1.1 \cdot 10^{15}$, care has to be taken when implementing this recursion, as it is not possible to store arbitrary subsets of I. Using the ideas put forward at the introduction of this section, it is possible to formulate the recursion in a way that at most 15 nodes are considered at the same time, corresponding to a 3×5 grids. The technical details are in the Supplementary Material Appendix F [Wang, Johannes and Massam (2018)].

To obtain an outer approximation F_2, we adapt Strategy 3.2 of Section 3.2 and cover the graph with 5×3 grid subgraphs. These subgrids are supported on the same vertex subsets $V_i, i = 1, \dots, 8$ as used when computing F_1. This makes it possible to compare F_1 and F_2. For $i = 1, 3, \dots, 8$, we compute $F_{2,V_i} = F_{\Delta|_{V_i}}(\pi_{V_i}(I_+))$. The outer approximation is then $F_2 = \bigcap_i \pi_{V_i}^{-1}(F_{2,V_i})$. Again, we do not compute F_2 explicitly, but we only store F_{2,V_i} in a computer as a representation of F_2.

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 3×5 subgrids, but only a cover of four 3×5 subgrids and one 2×5 subgrid. 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) F_t for high-dimensional models. To find an inner and an outer approximation to F_t, we divide

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%

the original problem into subproblems with at most 16 variables for which we can use linear programming. Then we combine the facial sets of the subproblems and relate them to F_t. Identifying the subproblems and relating the facial sets to F_t is numerically easy, and the corresponding software can be obtained upon request, from the authors.

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 parameters than 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 (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.

SUPPLEMENTARY MATERIAL

Supplement to "Approximating faces of marginal polytopes in discrete hierarchical models." (DOI: 10.1214/18-AOS1710SUPP; .pdf). 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 μ_L introduced in Section 2. Appendices D and E give further results for the examples from Section 5. Appendix F gives the technical details for the example in Section 6.2.

REFERENCES

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333

CSISZÁR, I. and MATÚŠ, F. (2008). Generalized maximum likelihood estimates for exponential families. *Probab. Theory Related Fields* **141** 213–246. MR2372970

CSISZÁR, I. and SHIELDS, P. (2004). *Information Theory and Statistics*: *A Tutorial*, 1st ed. Now Publishers, Hanover, MA.

DEZA, M. M. and LAURENT, M. (2010). *Geometry of Cuts and Metrics*. *Algorithms and Combinatorics* **15**. Springer, Heidelberg. MR2841334

DOBRA, A., EROSHEVA, E. A. and FIENBERG, S. E. (2004). Disclosure limitation methods based on bounds for large contingency tables with applications to disability. In *Statistical Data Mining and Knowledge Discovery* 93–116. Chapman & Hall, Boca Raton, FL. MR2048950

DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

ERIKSSON, N., FIENBERG, S. E., RINALDO, A. and SULLIVANT, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.* **41** 222–233. MR2197157

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, MA. MR0623082

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

GAWRILOW, E. and JOSWIG, M. (2000). Polymake: A framework for analyzing convex polytopes. In *Polytopes—Combinatorics and Computation* (*Oberwolfach*, 1997). *DMV Sem.* **29** 43–73. Birkhäuser, Basel. MR1785292

GEYER, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.* **3** 259–289. MR2495839

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. The Univ. Chicago Press, Chicago, IL. MR0408098

KARWA, V. and SLAVKOVIĆ, A. (2016). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.* **44** 87–112. MR3449763

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann. Statist.* **40** 861–890. MR2985936

LIU, Q. and IHLER, A. (2012). Distributed parameter estimation via pseudo-likelihood. *Int. Conf. Mach. Learn.* (*ICML*).

MASSAM, H. and WANG, N. (2015). A local approach to estimation in discrete loglinear models. Preprint. Available at arXiv:1504.05434.

MASSAM, H. and WANG, N. (2018). Local conditional and marginal approach to parameter estimation in discrete graphical models. *J. Multivariate Anal.* **164** 1–21. MR3738130

RAUH, J., KAHLE, T. and AY, N. (2011). Support sets in exponential families and oriented matroid theory. *Internat. J. Approx. Reason.* **52** 613–626. MR2787021

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

SCHMIDT, M. (2005). minFunc: Unconstrained differentiable multivariate optimization in Matlab. http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html.

VLACH, M. (1986). Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Discrete Appl. Math.* **13** 61–78. MR0829339

WANG, N., RAUH J. and MASSAM, H. (2019). Supplement to "Approximating faces of marginal polytopes in discrete hierarchical models." DOI:10.1214/18-AOS1710SUPP.

ZIEGLER, G. M. (1995). *Lectures on Polytopes. Graduate Texts in Mathematics* **152**. Springer, New York. MR1311028

N. WANG
H. MASSAM
DEPARTMENT OF MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO M3J 1P3
CANADA
E-MAIL: wangnanw@yorku.ca
massamh@yorku.ca

J. RAUH
MAX PLANCK INSTITUTE
FOR MATHEMATICS IN THE SCIENCES
INSELSTRAßE 21
04103 LEIPZIG
GERMANY
E-MAIL: jrauh@mis.mpg.de