Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIAL FOR "APPROXIMATING FACES OF MARGINAL POLYTOPES IN DISCRETE HIERARCHICAL MODELS"

By Nanwei Wang

York University, Toronto, Canada

AND

By Johannes Rauh

York University, Toronto, Canada

Max Planck Institute for Mathematics in the Sciences, Germany

AND

By Hélène Massam

York University, Toronto, Canada

APPENDIX A: PARAMETRIZING HIERARCHICAL MODELS

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

$$\log p(i) = \sum_{D \in \Delta} \theta_D(i_D) \tag{A1}$$

This parametrization is not identifiable; that is, for any joint distribution p from the hierarchical model there are different choices for the functions θ_D that satisfy (A1). One way to make the parameters unique is to choose a special element within each set I_v, which we denote by 0. The choice of 0 is arbitrary, and a different choice of 0 leads to a simple affine change of parameters. With this choice, the functions θ_D become unique if one requires $\theta_D(i_D) = 0$ whenever $i_v = 0$ for some $v \in D$.

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

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

For any $j \in J$, let

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

To simplify the notation, we write $j \lhd i$ whenever $S(j) \subseteq S(i)$ and $j_{S(j)} = i_{S(j)}$. It is convenient to introduce the vectors

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

where $e_j, j \in J$ are the unit vectors in R^J. Moreover, let A be the $J \times I$ matrix with columns f_i, $i \in I$, and let $\tilde{A}$ be the $(1 + |J|) \times I$ matrix with columns equal to $\binom{1}{f_i}$, $i \in I$. Then (2) can be rewritten in the following equivalent forms

$$\log p_\theta(i) = \sum_{j \in J : j \lhd i} \theta_j - k(\theta) = \langle \theta, f_i \rangle - k(\theta) = A^t\theta - k(\theta) = \tilde{A}^t\tilde{\theta}, \tag{A2}$$

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

$$-\theta_0 = k(\theta) = \log\Big(\sum_{i \in I} \exp\Big(\sum_{j \in J : j \lhd i} \theta_j\Big)\Big) \tag{A3}$$

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

$$\tilde{A}n = \begin{pmatrix} N \\ t \end{pmatrix} \quad \text{and} \quad An = t, \tag{A4}$$

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

It follows from (A4) that $\frac{t}{N} = \sum_{i \in I} \frac{n(i)}{N} f_i$. Therefore, t belongs to the convex polytope with extreme points $f_i, i \in I$. This polytope is the *marginal polytope* of the hierarchical model, denoted by $\mathbf{P}_\Delta$.

EXAMPLE A.1. Let $V = \{a, b, c\}$, $I_a = \{0, 1\} = I_b = I_c$ and $\Delta =$

$\{a, b, c, ab, bc\}$. Then

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

APPENDIX B: EXAMPLE: TWO BINARY RANDOM VARIABLES

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

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

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

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

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

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

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

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

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

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

Each line of the last column contains three combinations of the parameters $\theta_i^{(s)}$ that converge to a finite value. Any other parameter combination that converges is a linear combination of these three. This can be seen by using the coordinates μ_i introduced in Section 4.2. For example, on the facet $\mathbf{F}_{01}$, consider the parameters

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\text{(B5)} \quad 0 = \frac{\partial}{\partial \theta_{01}} l(\theta) &= -\frac{\exp(\theta_{01}^{(1)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})}{1 + \exp(\theta_{01}^{(1)}) + \exp(\theta_{10}^{(0)}) + \exp(\theta_{01}^{(1)} + \theta_{10}^{(0)} + \theta_{11}^{(0)})} \\
&= -\frac{2\exp(\theta_{01}^{(1)})}{1 + 2\exp(\theta_{01}^{(1)})}.
\end{aligned}$$

This derivative is negative for any finite value of $\theta_{01}^{(1)}$, and thus the critical equation has no finite solution. If we try to solve this equation numerically, we will find that $\theta_{01}^{(1)}$ will be a large negative number. Next, we look at θ_{10}. We fix the other variables and try to solve

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

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

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

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

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

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

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

For general data, the derivative of $l(\theta)$ with respect to θ_{01} (equation (B5)) takes the form

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

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

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

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

APPENDIX C: PARAMETRIZATIONS ADAPTED TO FACIAL SETS

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

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

(∗∗) $a^\mu_{j,i} = 0$ for all $i \in F_t$ and $j \in L \setminus L_t$.

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

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

$$\log p_{F_t,\mu}(i) = \langle \mu^t, f^\mu_i \rangle - k_F(\mu),$$
$$\log p_\mu(i) = \langle \tilde{\mu}^t, f^\mu_i \rangle - k(\mu),$$

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

$$(\log p_\mu)C = \mu_{L_t} \quad \text{and} \quad (\log p_{F_t,\mu})C = \mu_L. \tag{C6}$$

When $p_{\mu^{(s)}}$ is a sequence in $\mathcal{E}_A$ with limit p_μ in $\mathcal{E}_{F_t,A}$, then (C6) shows that $\mu^{(s)}_i \to \mu_i$ for $i \in L_t$. In this sense, we have remedied problem 2.

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

introduced in Remark 2.2, there exist c inequalities

$$\text{(C7)} \qquad \langle \tilde{g}_1, \tilde{x} \rangle \geq 0, \quad \ldots, \quad \langle \tilde{g}_c, \tilde{x} \rangle \geq 0$$

that together define $\mathbf{F}_t$. In this case, the vectors $\tilde{g}_1, \ldots, \tilde{g}_c$ are linearly independent and satisfy $\langle \tilde{g}_j, \tilde{f}_i \rangle = 0$ (thus, they are a basis of the kernel of $(\tilde{A}^{\mu_L}_{F_t})^t$). It follows that the kth component of g_j, denoted by $g_{j,k}$, vanishes if $k \in L_t$; that is, the inequalities (C7) do not involve the variables corresponding to L_t. Let G be the square matrix, indexed by $L \setminus L_t$ with entries $g_{j,k}$, $j, k \in L \setminus L_t$. Then the square matrix

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

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

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

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

This implies the following properties:

1. If all parameters λ_j with $j \notin L_t$ are sent to $-\infty$, then p_λ tends towards a limit distribution with support F_t.
2. The coefficient of λ_j in any log-probability is non-negative, so there is no cancellation of $\pm\infty$.

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

$$\langle \tilde{g}_j, \tilde{f}_{i_j} \rangle > 0, \qquad \text{and} \qquad \langle \tilde{g}_{j'}, \tilde{f}_{i_j} \rangle = 0 \text{ for all } j' \neq j,$$

and so

$$A^\lambda_{j,i_j} > 0, \qquad \text{and} \qquad A^\lambda_{j'}, i_j = 0 \text{ for all } j' \neq j.$$

Hence:

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

It is not difficult to see that, conversely, any parametrization that satisfies these three properties comes from facets defining the face $\mathbf{F}_t$.

APPENDIX D: UNIFORM SAMPLING FOR THE 4×4 GRID

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

sample size	MLE does not exist	$F_1 = F_t$	$F_2 = F_t$
10	98.5%	96.3%	100.0%
15	68.9%	99.9%	100.0%
20	29.0%	100.0%	100.0%
50	0.0%	100.0%	100.0%

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

APPENDIX E: ESTIMATED CELL FREQUENCIES FOR THE NLTCS DATA

The following table lists the estimates of the top five cell counts obtained using our method and compares them with those obtained by other methods in Dobra and Lenkoski (2011).

Support of Cell	Observed	GoM	LC	CGGMs	MLE on facial set
$\emptyset$	3853	3269	3836.01	3767.76	3647.4
$\{10\}$	1107	1010	1111.51	1145.86	1046.9
$\{1:16\}$	660	612	646.39	574.76	604.4
$\{5\}$	351	331	360.52	452.75	336
$\{5,10\}$	303	273	285.27	350.24	257.59
$\{12\}$	216	202	220.47	202.12	239.24

APPENDIX F: THE 5×10 GRID

This section presents the details for the example in Section 6.2. Let Δ be the simplicial complex of the 5×10 grid graph. We exploit the regularity of this graph and make use of the vertical separators in the grid to obtain inner and outer approximations of the facial sets. The graph has 50 nodes, which is too many to directly compute a facial set or even to store it. However,

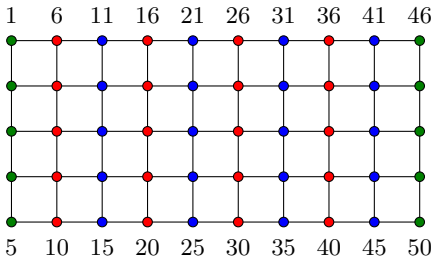

FIG 1. 5×10 *grid graph, the red and blue nodes are the set of separators we use to compute* F_1*, they are used iteratively to get a better lower approximation*

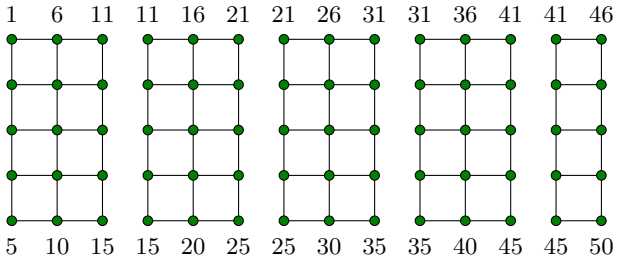

FIG 2. *Five induced subgrids*

the 5×10 grid has 8 vertical separators marked in red and blue in Figure 1, 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

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

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

Adding the blue separators to Δ gives a simplicial complex

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

with five irreducible components supported on the vertex sets V_1, V_3, V_5, V_7 and V_9 (Figure 3). To compute a facial set with respect to $\Delta_{S_2;S_4;S_6;S_8}$,

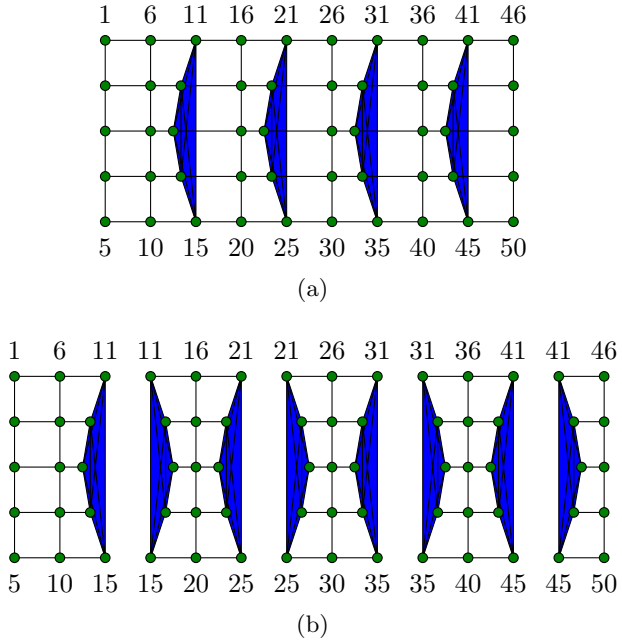

FIG 3. (a) *The* 5×10 *grid with the blue separators completed.* (b) *The five irreducible subcomplexes after completing the blue separators.*

according to Lemma 2.5 applied four times, we need to compute

$$G_{1,V_1} := F_{\Delta_{S_2}|_{V_1}}(\pi_{V_1}(I_+)), \quad G_{1,V_3} := F_{\Delta_{S_2;S_4}|_{V_3}}(\pi_{V_3}(I_+)),$$
$$G_{1,V_5} := F_{\Delta_{S_4;S_6}|_{V_5}}(\pi_{V_5}(I_+)), \quad G_{1,V_7} := F_{\Delta_{S_6;S_8}|_{V_7}}(\pi_{V_7}(I_+)),$$
$$G_{1,V_9} := F_{\Delta_{S_8}|_{V_9}}(\pi_{V_9}(I_+)).$$

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

We can improve the approximations by also considering the red separators

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

that separate $V_0 = \{1, \ldots, 10\}$, $V_2 = \{6, \ldots, 20\}$, $V_4 = \{16, \ldots, 30\}$, $V_6 = \{26, \ldots, 40\}$, $V_8 = \{36, \ldots, 50\}$. As explained in Section 3.1, we want to compute $G_1^{(2)} := F_{\Delta_{S_1;S_3;S_5;S_7}}(G_1)$. Again, instead of computing $G_1^{(2)}$ directly, we need only compute the much smaller sets $G_{1,V_0}^{(2)} := \pi_{V_0}(G_1^{(2)})$, $G_{1,V_2}^{(2)} := \pi_{V_2}(G_1^{(2)}), \ldots, G_{1,V_8}^{(2)} := \pi_{V_8}(G_1^{(2)})$. But is it possible to compute $G_{1,V_0}^{(2)}$, $G_{1,V_2}^{(2)}$, $\ldots$, $G_{1,V_8}^{(2)}$ from G_{1,V_1}, G_{1,V_3}, $\ldots$, G_{1,V_9}, without computing G_1 in between?

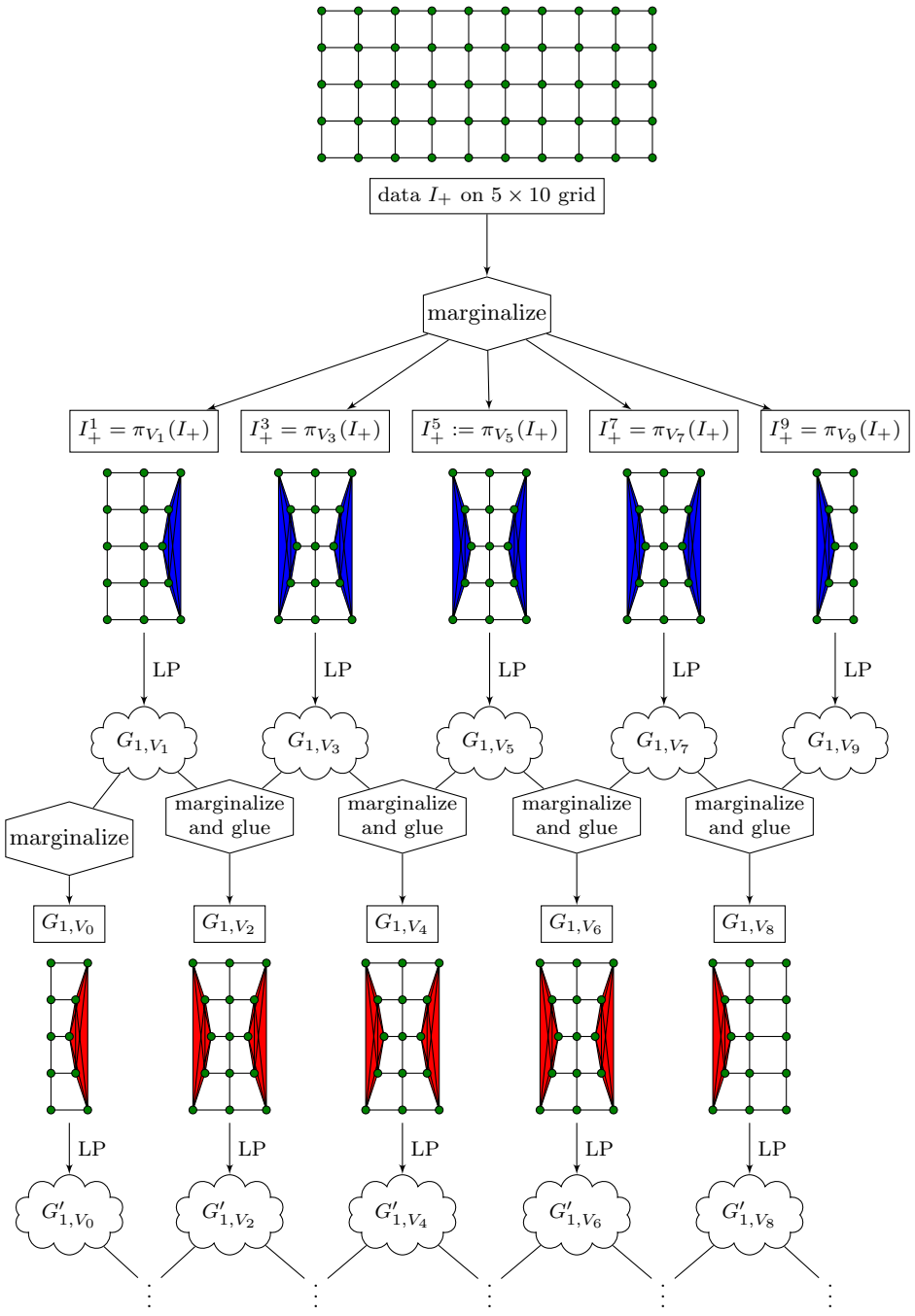

FIG 4. *Flow chart*

This is indeed the case: By Lemma 2.5, all we need to compute $G^{(2)}_{1,V_i}$ is $G_{1,V_j} := \pi_{V_j}(G_1)$, $j = i-1, i+1$. For $i = 0$, since $V_0 \subset V_1$, we can compute G_{1,V_0} from $\pi_{V_1}(G_1) = G_{1,V_1}$. For $i = 2, 4, 6, 8$, since $V_i \subset V_{i-1} \cup V_{i+1}$, we can compute G_{1,V_i} from $\pi_{V_{i-1}\cup V_{i+1}}(G_1)$, which itself can be obtained by "gluing" $\pi_{V_{i-1}}(G_1) = G_{1,V_{i-1}}$ and $\pi_{V_{i+1}}(G_1) = G_{1,V_{i+1}}$:

$$\pi_{V_{i-1}\cup V_{i+1}}(G_1) = \left(\pi^{V_{i-1}\cup V_{i+1}}_{V_{i-1}}\right)^{-1}(G_{1,V_{i-1}}) \cap \left(\pi^{V_{i-1}\cup V_{i+1}}_{V_{i+1}}\right)^{-1}(G_{1,V_{i+1}}),$$

where $\pi^{V'}_{V''}$ for $V'' \subseteq V'$ denotes the marginalization map from $I_{V'}$ to $I_{V''}$ and where $\left(\pi^{V'}_{V''}\right)^{-1}$ denotes the lifting from $I_{V''}$ to $I_{V'}$.

As explained in Section 3.1, this procedure can be iterated: From $G^{(2)}_1$ we want to compute $G^{(3)}_1 := F_{\Delta_{S_2;S_4;S_6;S_8}}(G'_1)$ or, more precisely, we want to compute $G^{(3)}_{1,V_i} = \pi_{V_i}(G^{(3)}_1)$ for $i = 1, 3, \ldots, 9$. Again, we do this without looking at $G^{(2)}_1$ directly by just using the information available through the $G^{(3)}_{1,V_i}$. Iterating this procedure, we obtain a sequence of sets $G^{(k)}_{1,V_i}, G^{(k)}_{1,V_j}$ (with odd i and even j), which stabilizes after a finite number of steps. Our best inner approximation is then $F_1 = \bigcap_{i=0}^{9} \pi^{-1}_{V_i}(F_{1,V_i})$, where $F_{1,V_i} := \bigcup G^{(k)}_{1,V_i}$. Again, we do not compute F_1 explicitly, but we represent it in terms of the F_{1,V_i}. The process is visualized in Figure 4.

Let us now consider the outer approximation F_2. We adapt Strategy 3 of Section 3.2 and cover the graph with 5×3 grid subgraphs, since the facial sets for such graphs can easily be computed. These subgrids are supported on the same vertex subsets $V_i, i = 1, \ldots, 8$ as used when computing F_1. This makes it possible to compare F_1 and F_2. For $i = 1, 3, \ldots, 8$ we compute $F_{2,V_i} = F_{\Delta|_{V_i}}(\pi_{V_i}(I_+))$. The outer approximation is then $F_2 = \bigcap_i \pi^{-1}_{V_i}(F_{2,V_i})$. Again, we don't compute F_2 explicitly, but we only store F_{2,V_i} in a computer as a representation of F_2. To compare the two approximations F_1 and F_2, we need only compare their projections F_{1,V_i} and F_{2,V_i} pairwise, $i = 1, \ldots, 8$.

REFERENCES

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

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

E-MAIL: wangnanw@yorku.ca E-MAIL: jarauh@yorku.ca

E-MAIL: massamh@yorku.ca