The maximum likelihood estimate in high-dimensional discrete graphical models

Hélène Massam

Department of Mathematics and Statistics
York University

July 10, 2017

Joint work with Johannes Rauh and Nanwei Wang

The problem

- We are given a multivariate r.v. $X = (X_v, v \in V)$ where V is a finite indexing set. Assume that the distribution of X is Markov w.r.t. an undirected graph $G = (V, E)$. Each node X_v takes values in a discrete set.
- The density of the distribution of X is of the form

$$f(x; \theta) = e^{\{\langle \theta, t(x) \rangle - k(\theta)\}}.$$

- Given the graph and the model, we want to compute the maximum likelihood estimate of the parameter θ.
- When computing the mle, we are faced to two major problems.
  - In high-dimensions, $k(\theta)$ is intractable and it is impossible to compute the mle.
  - In any dimension, if the data in the contingency table has zero counts, the mle might not exist. How do we know if the mle exists and if it does not, what should we do for inference?

Ingredients of a hierarchical model

- A discrete random variable $X = (X_v, v \in V)$, $\; x_v \in I_v = \{0, 1, \ldots, d_v\}$.
- Let $G = (V, E)$ an undirected graph. We have N sample points.
- A $p = |V|$-dimensional contingency table. The set of cells is

$$I = \prod_{v \in V} I_v = \{i = (i_1, \ldots, i_p), \;\; i_v \in I_v\}.$$

- The support of $i = (i_v, v \in V) \in I$ is

$$S(i) = \{v \in V \mid i_v \neq 0\}.$$

- Let Δ be a set of subsets of V such that if $D \in \Delta$ and $D_1 \subset D$, then $D_1 \in \Delta$.
- $J = \{i \in I \mid S(i) \in \Delta\} \subset I$. We use the **notation** $j \triangleleft i$

$$j \triangleleft i \iff S(j) \subset S(i) \;\text{ and }\; j_{S(j)} = i_{S(j)}.$$

The hierarchical model and the distribution of $n(i), i \in I$

- $\log p(i) = \theta_0 + \sum_{j \lhd i} \theta_j$, $\log p(0) = \theta_0$. (baseline constraints for θ_j).
- The multinomial distribution for the cell counts $(n(i), i \in I)$ is prop. to

$$\prod_{i \in I} p(i)^{n(i)} = \exp\{\langle \theta, t \rangle - k(\theta)\}, \quad \text{where}$$

$$\theta = (\theta_j, j \in J), \quad t = (t_j, j \in J), \quad t_j = \sum_{i : j \lhd i} n(i) = n(j_{s(j)}), \quad k(\theta) = \log \left(\sum_{i \in I} e^{\sum_{j \lhd i} \theta_j} \right)$$

- For each $i \in I$, we define the vector $f_i \in R^J$ by

$$(f_i)_j = \begin{cases} 1 & \text{if } j \lhd i \\ 0 & \text{otherwise} \end{cases}.$$

- Then we have

$$\begin{cases} L(\theta) \propto \exp\{\langle \theta, t \rangle - \log(\sum_{i \in I} e^{\langle \theta, f_i \rangle})\} & \text{the likelihood} \\ t = \sum_{i \in I} n(i) f_i & \text{the sufficient statistic} \end{cases}$$

Example: a graphical model

Let $V = \{a, b, c\}$, $I_a = \{0, 1\} = I_b = I_c$
G equal to $a \bullet$ ———— $\bullet\ b$ ———— $\bullet\ c$.
Then $\Delta = \{a, b, c, ab, bc\}$, , $J = \{(1, 0, 0),\ (0, 1, 0),\ (0, 0, 1),\ (1, 1, 0),\ (0, 1, 1)\}$,

$I = \{(0, 0, 0),\ (1, 0, 0),\ (0, 1, 0),\ (0, 0, 1),\ (1, 1, 0),\ (1, 0, 1),\ (0, 1, 1),\ (1, 1, 1)\}$,
$\theta = (\theta_{100}, \theta_{010}, \theta_{110}, \theta_{001}, \theta_{011})$.,
$t = (n(1++), n(+1+), n(++1), n(11+), n(+11))$

The f_i, $i \in I$ are

	f_{000}	f_{100}	f_{010}	f_{110}	f_{001}	f_{101}	f_{011}	f_{111}
a	0	1	0	1	0	1	0	1
b	0	0	1	1	0	0	1	1
ab	0	0	0	1	0	0	0	1
c	0	0	0	0	1	1	1	1
bc	0	0	0	0	0	0	1	1

The model is $(\log p(i)/p(000),\ i \in I \setminus \{(000)\}) = A^t\theta$ where A, the design matrix, is the matrix above and $\theta = (\theta_{100}, \theta_{010}, \theta_{110}, \theta_{001}, \theta_{011})$.

$$k(\theta) = \log\Big(1 + e^{\theta_{100}} + e^{\theta_{010}} + e^{\theta_{001}} + e^{\theta_{100}+\theta_{010}+\theta_{110}} + e^{\theta_{100}+\theta_{001}} + e^{\theta_{010}+\theta_{001}+\theta_{011}} + e^{\theta_{100}+\theta_{010}+\theta_{001}+\theta_{110}+\theta_{110}}\Big).$$

The polytope $\mathbf{P}_\Delta$ with extreme points $f_i, i \in I$ is called the marginal polytope of the model.

A simpler example with its geometric representation

Let $V = \{a, b\}, I_a = \{0, 1\} = I_b$ and let us consider the saturated model, that is the graphical model with graph G equal to $a \bullet \!\!\text{———}\!\! \bullet b$.
Then $\theta = (\theta_{10}, \theta_{01}, \theta_{11})$,
$\Delta = \{a, b, ab\}, \quad I = \{(0,0),\ (1,0),\ (0,1),\ (1,1)\} \quad J = \{(1,0),\ (0,1),\ (1,1)\}$.

The $f_i, i \in I$ are

	f_{00}	f_{10}	f_{01}	f_{11}
a	0	1	0	1
b	0	0	1	1
ab	0	0	0	1

The model is

$$\begin{pmatrix} \log p(10)/p(00) \\ \log p(01)/p(00) \\ \log p(11)/p(00) \end{pmatrix} = \begin{pmatrix} \theta_{10} \\ \theta_{01} \\ \theta_{10} + \theta_{01} + \theta_{11} \end{pmatrix} = \begin{pmatrix} f_{10}^t \\ f_{01}^t \\ f_{11}^t \end{pmatrix} \begin{pmatrix} \theta_{10} \\ \theta_{01} \\ \theta_{11} \end{pmatrix} = A^t\theta$$

A is the design matrix for the hierarchical model.

We also have $\tilde{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ with $\begin{pmatrix} \log p(00) \\ \log p(10) \\ \log p(01) \\ \log p(11) \end{pmatrix} = \tilde{A}^t \begin{pmatrix} \theta_{00} \\ \theta \end{pmatrix}$.

The marginal polytope

Two binary variables example, $V = \{a, b\}$

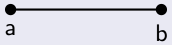


Figure 1: The simplicial complex $\Delta = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$

The facets are

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

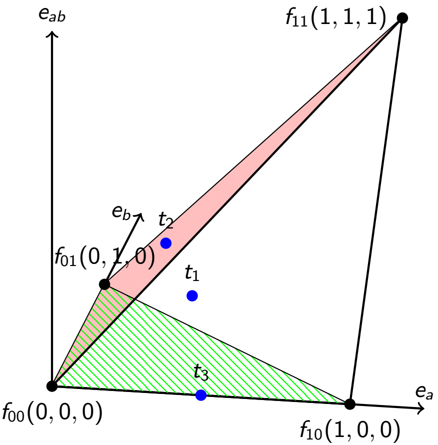


Figure 2: The marginal polytope

The maximum likelihood estimate $\hat{\theta}$ for this example

- The model is saturated. The solution to the likelihood equations is just the empirical distribution; that is, $p_*(i) = \frac{n(i)}{N}$.
- Suppose $t \in \mathbf{F}_{00}$ (i.e. $n = (0, n_{01}, n_{10}, n_{11})$ with $n(01), n(10), n(11) > 0$). The solution gives a sequence of values $\theta^{(s)}$ such that $p_{\theta^{(s)}}(00) \to 0$, while all other probabilities converge to a non-zero value. It follows that

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

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

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

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

The mle $\hat{\theta}$ does not exist but $p_*(i) = \frac{n(i)}{N}$ is well-defined.

Erroneous inference

- $\hat{\theta}$ does not exist but ... $\begin{cases} \theta^{(s)}_{01} + \theta^{(s)}_{00} = \log p_{\theta^{(s)}}(01) \\ \theta^{(s)}_{10} + \theta^{(s)}_{00} = \log p_{\theta^{(s)}}(10) \\ \theta^{(s)}_{11} + \theta^{(s)}_{01} = \log p_{\theta^{(s)}}(11)/p_{\theta^{(s)}}(10) \end{cases}$ all converge to a finite value.

  However the computer will give unreliable values of $\theta^{(s)}_j$ such that $\log p_{\theta^{(s)}}(i)$ do not converge to the right values.

- Moreover to test model M_1 vs. model M_2 where $d = \dim(M_1) - \dim(M_2)$, if t belongs to a face of the marginal polytope for at least one of the models, then the asymptotic distribution of

$$G^2 = 2\sum_{i\in I} n(i) \log \frac{\widehat{p^1(i)}}{\widehat{p^2(i)}} \not\to \chi^2_d.$$

  Indeed, instead $G^2 \to \chi^2_{d'}$ where $d' = d'_1 - d'_2$ with $d'_k, k = 1, 2$ being the dimension of the face $\mathbf{F}_k$ of M_k that t belongs to.

Conditions for the existence of the mle

Haberman (1974), Erikson et al. (2006), Fienberg and Rinaldo (2013): the mle exists iff the data vector t belongs to the interior of the marginal polytope with extreme points $f_i, i \in I$.

or equivalently

The mle does not exist iff the data vector belongs to a proper face (i.e. not the interior) of the marginal polytope.

We therefore have to identify the smallest face $\mathbf{F}_t$ of the marginal polytope of the model containing the data vector t

The facial set

- A face $\mathbf{F}$ of the marginal polytope $\mathbf{P}$ is identified by its facial set $F = \{i \in I \mid f_i \in F\}$.
- Since $t = \sum_{i \in I} n(i) f_i = \sum_{i \in I_+} n(i) f_i$, the facial set F_t of $\mathbf{F}_t$ is thus such that
$$F_t \supset I_+ = \{i \in I \mid n(i) > 0\}. \quad \text{crucial property}$$
- Recall $t \in R^J$ and so the hyperplane containing $\mathbf{F}_t$ will be defined by some $g \in R^{J+1}$ (one or more) such that
$$\langle g, \tilde{f}_i \rangle = 0, \; \forall i \in F_t,$$
- So, for sure if A_+ is the matrix with the columns indexed by I_+, and A_0 is the sub-matrix with columns indexed by $I \setminus I_+$ and g defines a supporting hyperplane, we have
$$g^t \tilde{A}_+ = 0 \;\text{ and }\; g^t \tilde{A}_0 \geqslant 0.$$
- Moreover, we want to find g such that, for all $f_i \notin F_t$, then $g^t f_i > 0$.

Linear Programming for Computing F_t (Fienberg & Rinaldo, 2012)

Lemma 1

Let A_+ and A_0 be as above. Solution g^* of the non-linear problem

$$\begin{array}{rl} \max_{g \in R^{J+1}} & z = \|g^t \tilde{A}\|_0 \\ s.t. & g^t \tilde{A}_+ = 0 \\ & g^t \tilde{A}_0 \geqslant 0 \end{array} \tag{1}$$

defines $\mathbf{F}_t$, the smallest face containing t. The corresponding facial set is $F_t = I \setminus supp(g^* A)$.

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

$$\begin{array}{rl} \max_{g \in R^{J+1}} & z = \|g^t \tilde{A}_0\|_1 \\ s.t. & g^t \tilde{A}_+ = 0 \\ & g^t \tilde{A}_0 \geq 0 \\ & g^t \tilde{A}_0 \leq 1 \end{array} \tag{2}$$

Here, we notice that only the support of data I_+ is needed to compute the facial set containing t, we don't need to know the exact cell counts.

Facial set approximation

- When p, the number of factors (or variables) is greater than 16, it is impossible to use linear programming to identify F_t.
  So, we will try to find approximations to F_t.

- **Definition**

  For the model generated by Δ and canonical statistic t, we define $F_\Delta(I_+)$ to be the smallest facial set containing I_+. Thus

$$F_t = F_\Delta(I_+).$$

- We use two principles for this approximation:
  1. reducibility of Δ
  2. If $\Delta' \subset \Delta$, then $F_t = F_\Delta(I_+) \subseteq F' = F_{\Delta'}(I_+)$

  Principle 2 above yields an inner and an outer approximation to F_t.
  outer approximation: If $\Delta_2 \subset \Delta$, then $F_t = F_\Delta(I_+) \subseteq F_2 = F_{\Delta_2}(I_+)$
  inner approximation:If $\Delta \subset \Delta_1$, then $F_1 = F_{\Delta_1}(I_+) \subseteq F_t = F_\Delta(I_+)$.

Reducible simplicial complex

Assume a simplicial complex Δ consists of some separable components, i.e. $\Delta = \Delta_1 \cup \Delta_2 \cup \ldots \cup \Delta_n$ and the separator $\Delta_{S_{ij}} = \Delta_i \cap \Delta_j$ is complete.

- any facet of some component $\mathbf{P}_{\Delta_i}$ is a facet of $\mathbf{P}_{\Delta}$. That is true because if $\Delta' \subset \Delta$, then f_i' is the projection of an f_i. Moreover, several f_i could be projected onto the same f_i'.
- any face of $\mathbf{P}_{\Delta}$ is either a face of a $\mathbf{P}_{\Delta_i}$ or the intersection of the faces of some components: this is true because if $\Delta_1 = \Delta_{|V_1}$ with $V_1 \subset V$, then each face of $\mathbf{P}_{\Delta_{|V_1}}$ corresponds to an inequality

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

  The same inequality also defines a face of $\mathbf{P}_{\Delta}$.
- $F_t = \cap_{i=1}^n F_{t_i}$ where t_i is the projection of t onto the model with simplicial complex Δ_i. **Erikson et al. (2006)**

If $\Delta' \subset \Delta$, then $F_\Delta(I_+) \subset F_{\Delta'}(I_+)$.

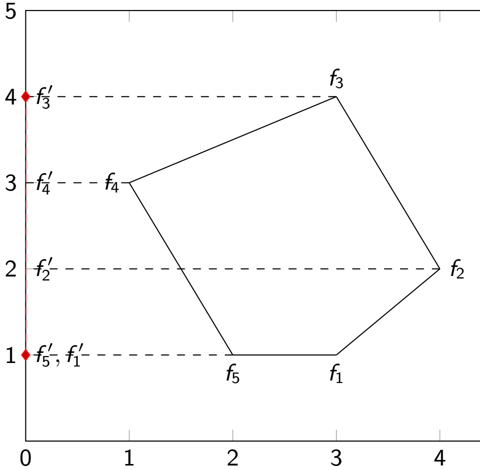
Let Δ, Δ' be two simplicial complexes with $\Delta' \subset \Delta$.

The polytope $\mathbf{P}_{\Delta'}$ is the projection of $\mathbf{P}_\Delta$ and the f_i' are the projections of f_i.
If $S = \{2, 3\}$, we see that

$$F_\Delta(S) = \{2, 3\},$$
$$F_{\Delta'}(S) = \{1, 2, 3, 4, 5\}.$$

We illustrated that if $\Delta' \subset \Delta$, for $S \subset I$, we have

$$F_\Delta(S) \subset F_{\Delta'}(S).$$



outer approximation

To get $F_2 \supset F_t$, we need $\Delta_2 \subset \Delta$.

For a large simplicial complex, we use sub-complexes defined by complete separators if they exist. We can prove that if no complete separators can be found, we can still work on a induced sub-simplicial complex.

1. Choose a subset of V: $a \subset V$. Apply LP on $\{\mathbf{P}_{\Delta_a}, I_{a+}\}$ and get a local facial set F_a,
2. F_a is a subset of I_a, we can extend F_a to a subset of I by adding all the configuration of $X_{V \setminus a}$: $F_2^1 = F_a \oplus I_{V \setminus a}$.

$$F_t \subseteq F_2^1,$$

3. Choose another subset of V: $b \subset V$, Repeat first two steps and get another outer approximation: $F_t \subseteq F_2^2$,
4. Improve the outer approximation by taking the intersection of all the outer approximation

$$F_t \subseteq \cap_i F_2^i$$

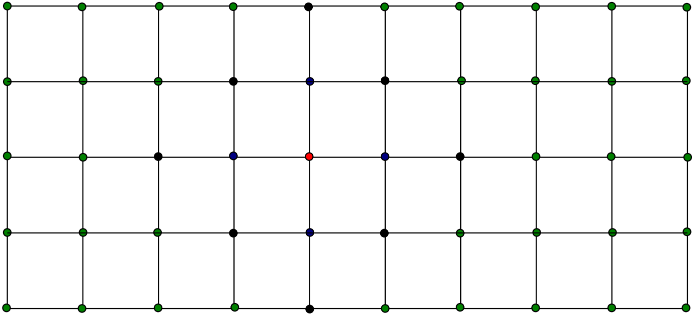
Inner approximation

To get $F_1 \subset F_t$, we need $\Delta_1 \supset \Delta$.

We can find and complete a proper separator to create a reducible simplicial complex.

1. Find and complete a separator set $S1$, apply LP to get a facial set F_1^1,
$$F_1^1 \subseteq F_t,$$
2. Use another separator set S_2, apply LP, but replace I_+ by F_1^1 to get another facial set F_1^2
$$F_1^1 \subseteq F_1^2 \subseteq F_t$$
3. Find other separator sets or repeat the first two steps iteratively, and we are getting closer and closer to the F_t:
$$F_1^1 \subseteq F_1^2 \subseteq \cdots \subseteq F_1^n \subseteq F_t$$

Model description



- 50 binary random variables and 135 parameters, 135×2^{50} design matrix,
- Sample from log-linear model whose parameters are randomly assigned as ± 0.5.

Outer approximation

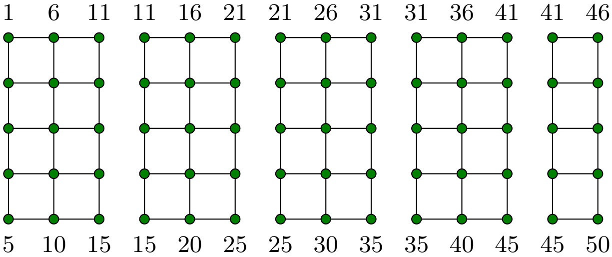


Figure 3: The 5 induced sub simplicial complexes for outer approximation

Inner approximation

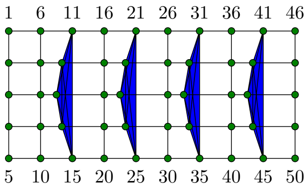


Figure 4: The 5 × 10 grid with blue separators completed

Figure 5: The 5 sub simplicial complexes after completing the separators

Applying two separator sets iteratively

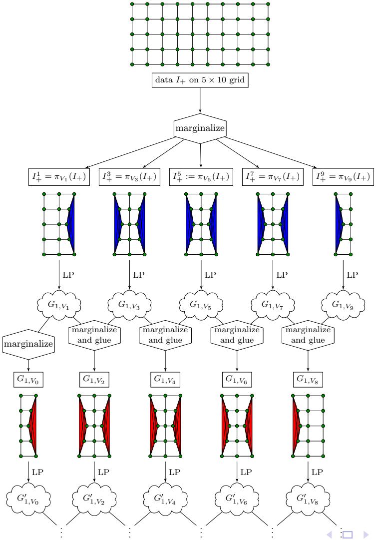


Figure 6: The flowchart of iterative steps

Numerical results

Table 1: 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%

Real data: data description and model selection

We consider the voting record of all 100 US senators on 309 bills from January 1 to November 19 2015. The votes are "yes" or "no".

- Dataset: 309 sample points of 100 binary random variables,
- We choose a model: we use the ℓ_1-regularized logistic regression to identify the neighbours of each variable and construct an Ising model. We set the penalty parameter to $\lambda = 32\sqrt{\log p/n} \approx 0.35$, resulting in a sparse graph.

Simplicial complex of the fitted model

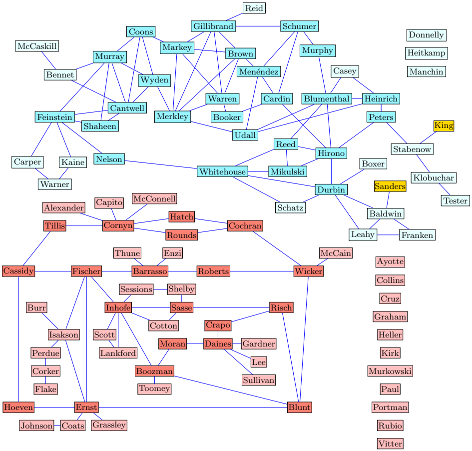


Figure 7: The golden nodes are independent senators, blue nodes are democratic and red nodes are republican.

Two prime components

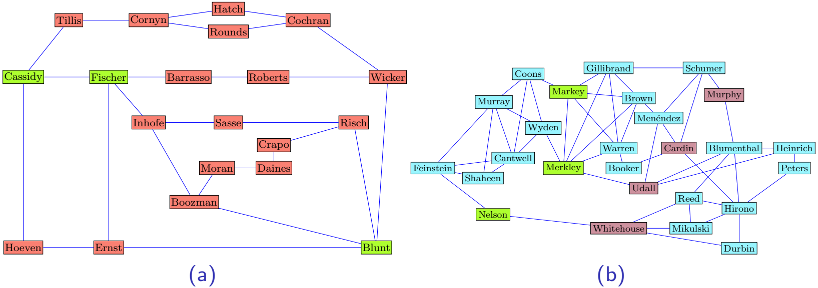


Figure 8: The simplicial complexes after cutting off the small prime components: (a) the republican party prime component Δ_r. (b) the democratic party prime component Δ_d. The yellow and pink nodes are the two separator sets we found to compute the approximation to the facial set.

Face computation of the republican party prime component

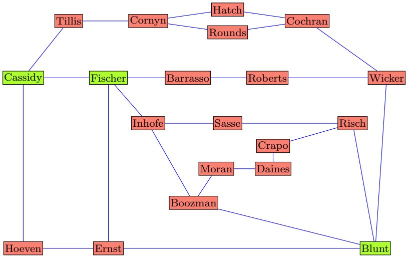


Figure 9: the republican party prime component Δ_r

- Δ_r includes 20 variables and 46 parameters, 46×2^{20} design matrix
- Choose separators: {"*Cassidy*","*Fisher*","*Blunt*"}
- Complete the separators, we will apply LP on two separable local simplicial complexes: $\Delta_{\tilde{\alpha}}$, $\Delta_{\tilde{\beta}}$,
- Both of the two local data $I_{\alpha+}, I_{\beta+}$ falls in the relative interior of the two marginal polytope $\mathbf{P}_{\Delta_{\tilde{\alpha}}}, \mathbf{P}_{\Delta_{\tilde{\beta}}}$,
- The original data I_r falls in the relative interior of the original marginal polytope $\mathbf{P}_{\Delta_r}$.

Face computation of the democratic party prime component

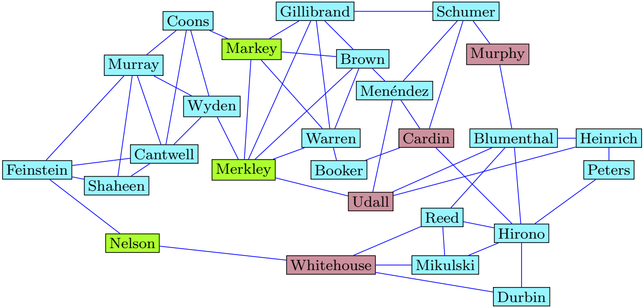


Figure 10: the democratic party prime component Δ_d

- Δ_d includes 26 variables and 77 parameters, 77×2^{26} design matrix,
- Separators: {"*Markey*","*Merkley*","*Nelson*"} and {"*Murphy*","*Cardin*","*Udall*","*Whitehouse*"}.

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

Table 2: Numbering of some senators

Outer approximation

Follow the separators, we choose three induced sub simplicial complexes from left to the right: Δ_α, Δ_β and Δ_γ, and split the original dataset into three local data $I_{\alpha+}$, $I_{\beta+}$ and $I_{\gamma+}$:

- Local data $I_{\alpha+}$ lies in the relative interior of $\mathbf{P}_{\Delta_\alpha}$
- Local data $I_{\beta+}$ lies on a face of $\mathbf{P}_{\Delta_\beta}$:

$$t_{warren} - t_{Gillibrand,warren} = 0,$$

- Local data $I_{\gamma+}$ lies on a face of $\mathbf{P}_{\Delta_\gamma}$:

$$t_{reed} - t_{reed,Hirono} = 0.$$

Therefore the outer approximation is the intersection of the above two faces:

$$\begin{cases} t_{warren} - t_{Gillibrand,warren} & = 0, \\ t_{reed} - t_{reed,Hirono} & = 0. \end{cases} \tag{3}$$

We denote this face as $\mathbf{F}_2$

Inner approximation

We complete the two separator sets respectively, and end up with three sub simplicial complexes from left to right: $\Delta_{\tilde{\alpha}}$, $\Delta_{\tilde{\beta}}$ and $\Delta_{\tilde{\gamma}}$, and the same local data $I_{\alpha+}$, $I_{\beta+}$ and $I_{\gamma+}$:

- Local data $I_{\alpha+}$ lies on a facet of $\mathbf{P}_{\Delta_{\tilde{\alpha}}}$:

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

- Local data $I_{\beta+}$ lies on a face of $\mathbf{P}_{\Delta_{\tilde{\beta}}}$:

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

- Local data $I_{\gamma+}$ lies on a face of $\mathbf{P}_{\Delta_{\tilde{\gamma}}}$:

$$\begin{cases} \langle g_7,\ t_{\tilde{\gamma}}\rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\ \langle g_8,\ t_{\tilde{\gamma}}\rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\ \langle g_9,\ t_{\tilde{\gamma}}\rangle = t_{23} - t_{23,53} = 0 \end{cases}.$$

Taking the intersection of faces of three simplicial complexes, we get the inner approximation:

$$\begin{cases} \langle g_1', \ t_{\tilde{d}} \rangle = t_{41} - t_{22,41} - t_{41,70} + t_{22,41,70} = 0 \\ \langle g_2', \ t_{\tilde{d}} \rangle = t_{87} - t_{56,87} = 0 \\ \langle g_3', \ t_{\tilde{d}} \rangle = t_{47,52,61} + t_{37,52} - t_{37,52,61} - t_{37,47,52} = 0 \\ \langle g_4', \ t_{\tilde{d}} \rangle = t_{37,47,52,61} - t_{47,52,61} = 0 \\ \langle g_5', \ t_{\tilde{d}} \rangle = t_{37,52} + t_{26} - t_{26,52} - t_{26,37} = 0 \\ \langle g_9', \ t_{\tilde{d}} \rangle = t_{23} - t_{23,53} = 0 \end{cases},$$

This is the smallest face of $\mathbf{P}_{\Delta_{\tilde{d}}}$ containing I_+. We denote it by $\mathbf{F}_{t_{\tilde{d}}}$, which is also the inner approximation $\mathbf{F}_1$.

Now we have $\mathbf{F}_1 \subset \mathbf{F}_2$, but $F_t \stackrel{?}{=} F_2$.
Observe:

$$\begin{aligned} \Delta_{\tilde{d}} &= \Delta_d + \text{added edges} \\ \mathbf{P}_{\Delta_{\tilde{d}}} &= \mathbf{P}_{\Delta_d} + \text{some more dimensions} \end{aligned}$$

same data l_+, but different sufficient statistics t_d and $t_{\tilde{d}}$.
Conclude:

- Any face of $\mathbf{P}_{\Delta_d}$ containing l_+ is also a face of $\mathbf{P}_{\Delta_{\tilde{d}}}$ containing l_+,

$$\langle g, t_d \rangle \geq c \Rightarrow \langle \tilde{g}, t_{\tilde{d}} \rangle \geq c, \text{where } \tilde{g} = [g, 0_{t_{\tilde{d}} \backslash t_d}]$$

- For any vector g that is perpendicular to $\mathbf{F}_{t_d}$, $\tilde{g}$ is perpendicular to $\mathbf{F}_{t_{\tilde{d}}}$. i.e.

$$\tilde{g} = k_1 g_1^{'} + k_2 g_2^{'} + k_3 g_3^{'} + k_4 g_4^{'} + k_5 g_5^{'} + k_6 g_6^{'}$$

- the values of k have to satisfy $k_1 = k_3 = k_4 = k_5 = 0$, since $t_{22,41,70}, t_{37,52,61}, t_{37,47,52,61}$ and $t_{37,52}$ are added dimensions,
- The equation of F_t can only be

$$\begin{cases} t_{87} - t_{56,87} = 0, \\ t_{23} - t_{23,53} = 0 \end{cases}$$

- $F_t = F_2$.

Now what?

Now that we have found the equations of the face containing t, how do we draw correct inference?

We want to write the exponential model on the face $\mathbf{F}_t$. To do so:

- we have the equation $\langle g_1, t\rangle = t_{87} - t_{56,87} = 0, \ \ \langle g_2, t\rangle = t_{23} - t_{23,53} = 0$. So, in principle we can identify all the $i \in I$ such that $\langle g_1, f_i\rangle = \langle g_2, f_i\rangle = 0$ i.e. all the $f_i \in F_t$ and build the new model $\log p = A^t_{new}\theta_{new}$ but there are many such $_i$'s.
- we use the parametrization $\mu_i = \log \frac{p(i)}{p(0)} = \langle \theta, f_i\rangle, i \in F_t \cap J$. These are identifiable and estimable parameters using the likelihood function

$$L(\mu) = \exp \sum_{j \in F_t \cap J} \mu_j \Big(\sum_{k \in F_t | j \lhd k} n(k)\Big) - N \log\Big(\sum_{i \in F_t} e^{\mu_i}\Big)$$

  where those $\mu_i, i \in F_t \setminus J$ are functions of $\mu_i, i \in F_t \cap J$. This is so because the $f_i, i \in F_t \setminus J$ are function of $f_i, i \in F_t \cap J$.
- the combinations of θ_j(in the old model) that are estimable are the $\langle \theta, f_i\rangle$, which, as we know, are equal to $\mu_i, i \in F_t \cap J$.

Now what? Continued

- when we compare two models for model selection, using the likelihood ratio statistic G^2 or the chi-square statistics χ^2, the degrees of freedom for the asymptotic distribution is the difference in the dimension of the faces containing the data vector in the two models.
- When we work with the parametrization $\mu_i, i \in F_t \cap J$, the matrix of second derivatives (i.e. the Hessian) estimated at the mle is nonsingular and we can give the usual confidence region for the parameter μ.