The geometry of discrete loglinear models and Bayes factors

Hélène Massam

York University

with G. Letac, Université Paul Sabatier

Overview of the problem

- N objects are classified according to $|V|$criteria and the data is gathered in a $|V|$-dimensional contingency table.
- For a given hierarchical loglinear model M, the cell counts follow a multinomial distribution with NEF form $f(y|\theta, M) = e^{\langle\theta,y\rangle - N k_\mu(\theta)}\mu(y), \quad \theta \in \Theta_M$.
- We put the Diaconis-Ylvisaker prior on θ, i.e.

$$\pi(\theta|\alpha, m, M)d\theta = \frac{1}{I(m,\alpha)} e^{\alpha\langle\theta,m\rangle - \alpha k_\mu(\theta)} d\theta$$

- The posterior dist. of θ is

$$\begin{aligned}\pi(\theta|\alpha, m, y, M) &= \frac{1}{I(\frac{\alpha m+y}{\alpha+N}, \alpha)} e^{\langle\theta,\alpha m+y\rangle - (\alpha+N)k_\mu(\theta)} \\ &= \frac{1}{I(\frac{\alpha m+y}{\alpha+N}, \alpha)} e^{(\alpha+N)\langle\theta,\frac{\alpha m+y}{\alpha+N}\rangle - (\alpha+N)k_\mu(\theta)}\end{aligned}$$

Overview of the problem

- If we compare two models M_1 and M_2, the Bayes factor is

$$B_{1,2} = \frac{I(\frac{\alpha m_1 + y_1}{\alpha + N}, \alpha + N)}{I(\frac{\alpha m_2 + y_2}{\alpha + N}, \alpha + N)} \frac{I(m_2, \alpha)}{I(m_1, \alpha)}.$$

Statistical folklore says that as $\alpha \to 0$, the Bayes factor favours the sparser model.

- For $M_1 =$ the saturated model on factors a, b, c and $M_2 =$ the decomposable model with interaction ab, bc, Steck and Jaakkola (2002) showed in fact that was not always the case and that the behaviour of the Bayes factor in fact depended upon the data.
- This has lead us to examine the behaviour of $I(m, \alpha)$ and $I(\frac{\alpha m + y}{\alpha + N}, \alpha + N)$ when $\alpha \to 0$

The data in a contingency table

- N objects are classified according to $|V|$criteria.
- We observe the value of $X = (X_\gamma |\ \gamma \in V)$ which takes its values (or levels) in the finite set I_γ.
- The data is gathered in a $|V|$-dimensional contingency table with

$$|I| = \times_{\gamma \in V} |I_\gamma| \ \text{ cells } \ i.$$

- The cell counts $(n) = (n(i), i \in \mathcal{I})$ follow a multinomial $\mathcal{M}(N, p(i), i \in \mathcal{I})$distribution.
- We denote $i_E = (i_\gamma, \gamma \in E)$ and $n(i_E)$ respectively the marginal-E cell and cell count.

The hierarchical loglinear model

• The generating set of the model is $\mathcal{D}$, a set of subsets of V such that if $D \in \mathcal{D}$ and $D_1 \subset D$, then $D_1 \in \mathcal{D}$. (also called hypergraph or abstract simplicial complex)

• Let $\Omega_{\mathcal{D}}$ the linear subspace of $x \in \mathbb{R}^I$ such that there exist functions $\lambda_D \in \mathbb{R}^I$ for $D \in \mathcal{D}$ depending only on i_D and such that $x = \sum_{D \in \mathcal{D}} \lambda_D$, that is

$$\Omega_{\mathcal{D}} = \{x \in \mathbb{R}^I : \exists \lambda_D \in \mathbb{R}^I, D \in \mathcal{D} \text{ such that } \lambda_D(i) = \lambda_D(i_D) \text{ and } x = \sum_{D \in \mathcal{D}} \lambda_D\}$$

The hierarchical model generated by $\mathcal{D}$ is the set of probabilities $p = (p(i))_{i \in I}$ on I such that $p(i) > 0$ for all i and such that $\log p \in \Omega_{\mathcal{D}}$. It is convenient to write for p in $\Omega_{\mathcal{D}}$

$$\log p(i) = \lambda_\emptyset + \sum_{D \in \mathcal{D}} \lambda_D(i) \tag{1}$$

where $\lambda_\emptyset$ does not depend on i and is thus a constant. The representation (1) is not unique.

The parametrization

- We choose a special cell $0 = (0, \ldots, 0)$.
- The generating set is $\mathcal{D} = \{D \subseteq V : \ D_1 \subset D \Rightarrow D_1 \in \mathcal{D}\}$.
- We write $S(i) = \{\gamma \in V : \ i_\gamma \neq 0\}$ and

$$j \lhd i \ \text{ if } \ S(j) \subseteq S(i) \text{ and } j_{S(j)} = i_{S(j)}.$$

- The parametrization: $p(i) \mapsto \theta_i = \sum_{j \lhd i} (-1)^{|S(i) \setminus S(j)|} \log p(j)$.
- Define

$$\begin{aligned} J &= \{j \in I : \ S(j) \in \mathcal{D}\} \\ J_i &= \{j \in J, \ j \lhd i\} \end{aligned}$$

- Then the hierarchical loglinear model can be written as

$$\log p(i) = \theta_\emptyset + \sum_{j \in J_i} \theta_j \quad \text{with} \quad \log p(0) = \theta_\emptyset.$$

The multinomial hierarchical model

$p(0) = e^{\theta_0} = (1 + \sum_{i \in I \setminus \{0\}} \exp \sum_{j \in J_i} \theta_j)^{-1} = L(\theta)^{-1}$ and

$$\prod_{i \in I} p(i)^{n(i)} = \frac{1}{L(\theta)^N} \exp\{\sum_{j \in J} n(j_{S(j)})\theta_j\} = \exp\{\sum_{j \in J} n(j_{S(j)})\theta_j + N\theta_0\}.$$

Then $\prod_{i \in I} p(i)^{n(i)}$ becomes

$$\begin{aligned} f(t_J|\theta_J) &= \exp\left\{\sum_{j \in J} n(j_{S(j)})\theta_j - N \log(1 + \sum_{i \in I \setminus \{0\}} \exp \sum_{j \in J_i} \theta_j)\right\} \\ &= \frac{\exp\langle \theta_J, t_J \rangle}{L(\theta_J)^N} = e^{\langle \theta_J, t_J \rangle - Nk(\theta_J)} \end{aligned}$$

with $\theta_J = (\theta_j, j \in J)$, $t_J = (n(j_{S(j)}), j \in J)$ and
$L(\theta_J) = (1 + \sum_{i \in I \setminus \{0\}} \exp \sum_{j \in J_i} \theta_j) = \sum_{i \in \mathcal{I}} \exp \sum_{j \in J_i} \theta_j$.

The measure generating the multinomial

Let $(e_j, j \in J)$ be the canonical basis of R^J and let

$f_i = \sum_{j \in J, j \lhd i} e_j, \quad i \in I.$ For G : $\overset{a}{\bullet} - \overset{b}{\bullet} - \overset{c}{\bullet}$ with binary data, $\mathcal{D} = \{a, b, c, ab, bc\}$ and $J = \{(100), (010), (001), (110), (011)\}$

$\mathcal{D}$	f_0	f_a	f_b	f_c	f_{ab}	f_{ac}	f_{bc}	f_{abc}
e_a	0	1	0	0	1	1	0	1
e_b	0	0	1	0	1	0	1	1
e_c	0	0	0	1	0	1	1	1
e_{ab}	0	0	0	0	1	0	0	1
e_{bc}	0	0	0	0	0	0	1	1

Here $R^I = R^8$ while $R^J = R^5$.

The Laplace transform of $\mu_J = \sum_{i \in \mathcal{I}} \delta_{f_i}$ is, for $\theta \in R^J$,

$$\int_{R^J} e^{\langle \theta, x \rangle} \mu_J(dx) = 1 + \sum_{i \in \mathcal{I} \setminus \{0\}} e^{\langle \theta, f_i \rangle} = 1 + \sum_{i \in \mathcal{I} \setminus \{0\}} e^{\sum_{j \lhd i} \theta_j} = L(\theta).$$

The DY conjugate prior

Therefore the multinomial $f(t_J|\theta_J) = \frac{\exp\langle\theta_J, t_J\rangle}{L(\theta_J)^N}$ is the NEF generated by μ_J^{*N}.

C_J is the open convex hull of the support of μ_J;
$f_i, i \in I$ are the extreme points

The Diaconis and Ylvisaker (1974) conjugate prior for θ

$$\pi(\theta_J|m_J, \alpha) = \frac{1}{I(m_J, \alpha)} e^{\{\alpha\langle\theta_J, m_J\rangle - \alpha \log L(\theta_J)\}}$$

is proper when the hyperparameters $m_J \in C_J$ and $\alpha > 0$.

Interpretation of the hyper parameter $(\alpha m_J, \alpha)$:

- α is the fictive total sample size
- $\alpha(m_j, \quad j \in J)$ represent the fictive marginal counts .

The Bayes factor between two models

The posterior density of J given t_J is

$$h(J|t_J) \propto \frac{I(\frac{t_J+\alpha m_J}{\alpha+N}, \alpha+N)}{I(m_J, \alpha)}.$$

Consider two hierarchical models defined by J_1 and J_2. The Bayes factor is

$$B_{1,2} = \frac{I(m_2, \alpha)}{I(m_1, \alpha)} \times \frac{I(\frac{t_1+\alpha m_1}{\alpha+N}, \alpha+N)}{I(\frac{t_2+\alpha m_2}{\alpha+N}, \alpha+N)}.$$

We will consider two cases depending on whether $\frac{t_k}{N} \in C_k,\ k = 1, 2$ or not.

The Bayes factor between two models

When $\alpha \to 0$,

- if $\frac{t_k}{N} \in C_k,\ k = 1, 2$, then

$$\frac{I(\frac{t_1+\alpha m_1}{\alpha+N}, \alpha + N)}{I(\frac{t_2+\alpha m_2}{\alpha+N}, \alpha + N)} \to \frac{I(\frac{t_1}{N}, N)}{I(\frac{t_2}{N}, N)}$$

which is finite. Therefore we only need to worry about $\lim_{\alpha\to 0} \frac{I(m_2,\alpha)}{I(m_1,\alpha)}$.

- if $\frac{t_k}{N} \in \bar{C}_k \setminus C_k,\ k = 1, 2$, then, we have to worry about $\lim_{\alpha\to 0} \frac{I(m_2,\alpha)}{I(m_1,\alpha)}$ and $\lim_{\alpha\to 0} \frac{I(\frac{t_1+\alpha m_1}{\alpha+N},\alpha+N)}{I(\frac{t_2+\alpha m_2}{\alpha+N},\alpha+N)}$.

The characteristic function of C

Definitions. Assume C is an open nonempty convex set in R^n.

- The support function of C is $h_C(\theta) = \sup\{\langle \theta, x\rangle \ : \ x \in C\}$
- The characteristic function of C:

$J_C(m) = \int_{R^n} e^{\langle \theta, m\rangle - h_C(\theta)} d\theta$

Examples of $J_C(m)$

- $C = (0,1)$. Then $h_C(\theta) = \theta$ if $\theta > 0$ and $h_C(\theta) = 0$ if $\theta \leq 0$.

Therefore $h_C(\theta) = \max(0, \theta)$ and

$$J_C(m) = \int_{-\infty}^{0} e^{\theta m} d\theta + \int_{0}^{+\infty} e^{\theta m - \theta} d\theta = \frac{1}{m(1-m)}.$$

Examples of $J_C(m)$

Examples of $J_C(m)$

- C is the simplex spanned by the origin and the canonical basis $\{e_1, \ldots, e_n\}$ in R^n and $m = \sum_{i=1}^n m_i e_i \in C$. Then

$$J_C(m) = \frac{n!\text{Vol}(C)}{\prod_{j=0}^n m_i} = \frac{1}{\prod_{j=1}^n m_i (1 - \sum_{j=1}^n m_i)}.$$

It can be computed as the limit of $\alpha^n I(m, \alpha)$ with

$$I(m,\alpha) = \int_{R^n} \frac{e^{\alpha\langle\theta,m\rangle}}{(1+\sum_{i=1}^n e^{\theta_i})^\alpha} d\theta = \int_{R^n} \frac{\prod_{i=1}^n e^{\alpha m_i \theta_i}}{(1+\sum_{i=1}^n e^{\theta_i})^\alpha} \prod_{i=1}^n d\theta_i = \frac{\prod_{i=0}^n \Gamma(\alpha m_i)}{\Gamma(\sum_{i=0}^n \alpha m_i)}$$

Examples of $J_C(m)$

- For the decomposable model with $\mathcal{D} = \{a, b, c, ab, bc\}$ with binary variables, we have

$$J = \{(1,0,0), (0,1,0), (0,0,1), (1,1,0), (0,1,1)\}$$

with C spanned by $f_j, j \in J$ and $m = \sum_{j \in J} m_j f_j$. Then, we obtain $J_C(m)$ as the limit of $\alpha^n I(m, \alpha)$

$$\begin{aligned} J_C(m) &= \frac{m_{(0,1,0)}(1 - m_{(0,1,0)})}{D_{ab} D_{bc}} \\ D_{ab} &= m_{(1,1,0)}(m_{(1,0,0)} - m_{(1,1,0)})(m_{(0,1,0)} - m_{(1,1,0)})(1 - m_{(1,0,0)} - m_{(0,1,0)} + m_{(1,1,0)}) \\ D_{bc} &= m_{(0,1,1)}(m_{(0,0,1)} - m_{(0,1,1)})(m_{(0,1,0)} - m_{(0,1,1)})(1 - m_{(0,0,1)} - m_{(0,1,0)} + m_{(0,1,1)}) \end{aligned}$$

Limiting behaviour of $I(m, \alpha)$

Theorem

Let μ be a measure on R^n,$n = |J|$, such that C the interior of the convex hull of the support of μ is nonempty and bounded. Let $m \in C$ and for $\alpha > 0$, let

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

Then

$$\lim_{\alpha \to 0} \alpha^n I(m, \alpha) = J_C(m).$$

Furthermore $J_C(m)$ is finite if $m \in C$.

Outline of the proof

$$
\begin{aligned}
I(m,\alpha) &= \int_{R^n} \frac{e^{\alpha\langle\theta,m\rangle}}{L(\theta)^\alpha} d\theta \\
\alpha^n I(m,\alpha) &= \int_{R^n} \frac{e^{\langle y,m\rangle}}{L(\frac{y}{\alpha})^\alpha} dy \quad \text{by chg. var. } y = \alpha\theta \\
L(\frac{y}{\alpha})^\alpha &= [\int_S e^{\frac{1}{\alpha}\langle y,x\rangle} \mu(dx)]^\alpha \\
&= \int_S [e^{\langle y,x\rangle}]^p \mu(dx) \Big)^{1/p} \quad \text{for } \alpha = 1/p, S = supp(\mu) \\
&= ||e^{\langle y,\bullet\rangle}||_p \to ||e^{\langle y,\bullet\rangle}||_\infty \quad \text{as } \alpha \to 0 \\
&= \sup_{x\in S} e^{\langle y,x\rangle} = \sup_{x\in C} e^{\langle y,x\rangle} = e^{\sup_{x\in C}\langle y,x\rangle}, \quad C = c.hull(S) \\
\alpha^n I(m,\alpha) &\to \int_{R^n} e^{\langle y,m\rangle - h_C(y)} dy = J_C(m)
\end{aligned}
$$

Limit of the Bayes factor

Let models J_1 and J_2 be such that $|J_1| > |J_2|$ and the data are in $C_i, ß = 1, 2$. Then the Bayes factor

$$\frac{I(m_2,\alpha)}{I(m_1,\alpha)} \frac{I(\frac{t_1+\alpha m_1}{\alpha+N}, \alpha+N)}{I(\frac{t_2+\alpha m_2}{\alpha+N}, \alpha+N)} \sim \alpha^{|J_1|-|J_2|} \frac{I(\frac{t_1}{N}, N)}{I(\frac{t_2}{N}, N)}$$

Therefore the Bayes factor tends towards 0, which indicates that the model J_2 is preferable to model J_1.

We proved the heuristically known fact that taking α small favours the sparser model.

We can say that α close to "0 " regularizes the model.

Important properties

We define the polar convex set C^o of C

$$C^o = \{\theta \in R^n \ ; \ \langle \theta, x \rangle \leq 1 \ \ \forall x \in C\}$$

then

- $\frac{J_C(m)}{n!} = \text{Vol}(C-m)^0 = \int_{C^o} \frac{d\theta}{(1-\langle \theta, m \rangle)^{n+1}}$

For the second equality, make the change of variable $\theta = \theta'/(1+\langle \theta', m \rangle)$

- If C in R^n is defined by its K $(n-1)$-dimensional faces $\{x \in R^n : \langle \theta_k, x \rangle = c_k\}$, then for $D(m) = \prod_{k=1}^K (\langle \theta_k, x \rangle - c_k)$,

$$D(m) J_C(m) = N(m)$$

where degree of $N(m)$ is $\leq K$.

Limiting behaviour of $I(\frac{\alpha m+t}{\alpha+N}, \alpha + N)$

We now consider the case when $\frac{t}{N} \in \overline{C} \setminus C$.
We write $\frac{\alpha m+t}{\alpha+N} = \lambda m + (1-\lambda)\frac{t}{N}$ with $\lambda = \frac{\alpha}{\alpha+N}$.

First step: Prove that when $\alpha \to 0$ i.e. $\lambda \to 0$ and $\frac{t}{N}$ belongs to a face of C of dimension k, then

$$\lim \lambda^{|J|-k} J_C(\lambda m + (1-\lambda)\frac{t}{N})$$

exist and is positive.

Second step: Show that $\lim \lambda^{|J|-k} D(\lambda)$ exist and is positive with

$$D(\lambda) = \mathbb{J}_C(\lambda m + (1-\lambda)y) - (\frac{N}{1-\lambda})^n I(\lambda m + (1-\lambda)y, \frac{N}{1-\lambda})$$

Limiting behaviour of $I(\frac{\alpha m+t}{\alpha+N}, \alpha+N)$

This will prove that

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

exists and is positive and therefore

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

where D is a positive constant.

Outline of the proof of

$$\lim_{\lambda \to 0} \lambda^{|J|-k} J_C(\lambda m + (1-\lambda)\frac{t}{N})$$

where we note $m = 0$ and $\frac{t}{N} = y$

- $$\frac{J_C((1-\lambda)y)}{n!} = \text{Vol}(C-(1-\lambda)y)^0 = \int_{C^o} \frac{d\theta}{(1-(1-\lambda)\langle\theta, y\rangle)^{n+1}}$$
- Parametrize C^o: consider the face F of C containing y. The dual face $\widehat{F}$ of C^o is

$$\widehat{F} = \{\theta \in \overline{C^o} \,|\langle\theta, f\rangle = 1 \;\forall f \in \mathcal{I}\} = \{\theta \in C^o \,|\langle\theta, y\rangle = 1\}.$$

- Cut $\overline{C^o}$ into "slices" $\widehat{F}_\epsilon = \{\theta \in \overline{C^o}\;;\; \langle\theta, y\rangle = 1-\epsilon\}$ and show $\text{vol}_{n-1}\widehat{F}_\epsilon \sim c\epsilon^k$
- $$\int_{\overline{C^o}} \frac{d\theta}{(1-(1-\lambda)\langle\theta, y\rangle)^{n+1}} = \int_0^\infty \frac{\text{vol}_{n-1}\widehat{F}_\epsilon d\epsilon}{(1-(1-\lambda)(1-\epsilon))^{n+1}} = \int_0^\infty \frac{f(\epsilon)d\epsilon}{(1-(1-\lambda)(1-\epsilon))^{n+1}}$$
- Using $f(\epsilon) \sim c\,\epsilon^k$ we will now show that $\lim_{\lambda\to 0} \lambda^{n-k} \int_0^\infty \frac{f(\epsilon)d\epsilon}{(1-(1-\lambda)(1-\epsilon))^{n+1}} = c\,B(k+1, n-k)$, and this concludes the proof.

Some facets of C

Let $\mathcal{D}$ be the generating set of the hierarchical model.

For each $D \in \mathcal{D}$ and each $j_0 \in J$ such that $S(j_0) \subset D$ define the affine forms

$$\begin{aligned} g_{0,D}(t) &= 1 + \langle g_{0,D}, t \rangle \\ g_{j_0,D}(t) &= \langle g_{j_0,D}, t \rangle. \end{aligned}$$

where

$$\begin{aligned} g_{0,D} &= \sum_{j; S(j) \subset D} (-1)^{|S(j)|} e_j \\ g_{j_0,D} &= \sum_{j; S(j) \subset D,\ j_0 \lhd j} (-1)^{|S(j)|-|S(j_0)|} e_j \end{aligned}$$

Facets of C common to all models

All subsets of the form

$$F(j,D) = H(j,D) \cap \overline{C}$$

with $H(j,D) = \{t \in \mathsf{R}^J \ ; \ g_{j,D}(t) = 0\}, \ D \in \mathcal{C}, \ S(j) \subset D$

$\mathcal{C} = \{$maximal elements of $\mathcal{D}\}$, are facets of C.

Example $a - - - b - - - c$. The facets are

$$t_{ab} = 0, \ t_a - t_{ab} = 0, \ t_b - t_{ab} = 0, \ 1 - t_a - t_b + t_{ab} = 0$$

and

$$t_{bc} = 0, \ t_b - t_{bc} = 0, \ t_c - t_{bc} = 0, \ 1 - t_b - t_c + t_{bc} = 0.$$

The facets of C when G is decomposable

For decomposable models,

$$H(j,D) = \{m \in \mathsf{R}^J \; ; \; g_{j,D}(m) = 0\}, \; D \in \mathcal{C}, \; S(j) \subset D$$

are the only facets of C.

Example $a - - - b - - - c$. The facets are

$D = ab$	$D = bc$
$t_{ab} = 0, j = (1,1,0);$	$t_{bc} = 0, j = (0,1,1);$
$t_b - t_{ab} = 0, j = (0,1,0);$	$t_c - t_{bc} = 0, j = (0,0,1);$
$t_a - t_{ab} = 0, j = (1,0,0)$	$t_b - t_{bc} = 0, j = (0,1,0)$
$1 - t_a - t_b + t_{ab} = 0, \; S(j) = \emptyset$	$1 - t_b - t_c + t_{bc} = 0, \; S(j) = \emptyset.$

The facets: traditional notation

Example $a---b---c$. For binary data, the facets are

$$
\begin{aligned}
Nt_{ab} = 0 &= n_{11+} \\
N(t_a - t_{ab}) = 0 &= n_{1++} - n_{11+} = n_{10+} \\
N(t_b - t_{ab}) = 0 &= n_{+1+} - n_{11+} = n_{01+} \\
N(1 - t_a - t_b + t_{ab}) = 0 &= N - n_{1++} - n_{+1+} + n_{11+} = n_{00+} \\
Nt_{bc} = 0 &= n_{+11} \\
N(t_b - t_{bc}) = 0 &= n_{+10} \\
N(t_c - t_{bc}) = 0 &= n_{+01} \\
N(1 - t_b - t_c + t_{bc}) = 0 &= n_{+00}
\end{aligned}
$$

The facets: traditional notation

Example: The complete model. Then $\mathcal{C} = \{abc\}$ and the facets are

$$
\begin{aligned}
Nt_{abc} = 0 &= n_{111} \\
N(t_{ab} - t_{abc}) = 0 &= n_{110} \\
N(t_{bc} - t_{abc}) = 0 &= n_{011} \\
N(t_{ac} - t_{abc} = 0 &= n_{101} \\
N(t_a - t_{ab} - t_{ac} + t_{abc}) = 0 &= n_{100} \\
N(t_b - t_{ab} - t_{bc} + t_{abc}) = 0 &= n_{010} \\
N(t_c - t_{ac} - t_{bc} + t_{abc}) = 0 &= n_{001} \\
N(1 - t_a - t_b - t_c + t_{ab} + t_{bc} + t_{ac} - t_{abc} = 0 &= n_{000}
\end{aligned}
$$

Steck and Jaakola (2002)

Steck and Jaakola (2002) considered the problem of the limit of the Bayes factor when $\alpha \to 0$ for Bayesian networks.

Bayesian networks are not hierarchical models but in some cases, they are Markov equivalent to undirected graphical models which are hierarchical models.

Problem: compare two models which differ by one directed edge only.

Equivalent problem: with three variables binary X_a, X_b, X_c each taking values in $\{0, 1\}$, compare
Model $\mathcal{M}_1$: $a - - - - b - - - - c$: $|J_1| = 5$.

Model $\mathcal{M}_2$: the complete model i.e. with $\mathcal{A} = \{(a, b, c)\}$.
$|J_2| = 7$

Generalization of S&J (2002)

They define

$$d_{EDF} = \sum_{i \in \mathcal{I}} \delta(n(i)) - \sum_{i_{ab} \in \mathcal{I}_{ab}} \delta(n(i_{ab})) - \sum_{i_{bc} \in \mathcal{I}_{bc}} \delta(n_{i_{bc}}) + \sum_{i_b \in \mathcal{I}_b} \delta(n(i_b))$$

where $\delta(x) = 0$ if $x = 0$ and $\delta(x) = 1$ otherwise. They show

$$\lim_{\alpha \to 0} B_{1,2} = \begin{cases} 0 & \text{if } d_{EDF} > 0 \\ +\infty & \text{if } d_{EDF} < 0 \end{cases}$$

We show that $d_{EDF} = k_1 - k_2$ and more generally if $\mathcal{C}_i$ and $\mathcal{S}_i$ the set of cliques and separators of the decomposable model $J_i,\ i = 1, 2$. We define

$$d_{EDF} = \sum_{C \in \mathcal{C}_1} \sum_{i_C \in \mathcal{I}_C} \delta(n(i_C)) - \sum_{S \in \mathcal{S}_1} \sum_{i_S \in \mathcal{I}_S} \delta(n(i_S)) - \Big(\sum_{C \in \mathcal{C}_2} \sum_{i_C \in \mathcal{I}_C} \delta(n(i_C)) - \sum_{S \in \mathcal{S}_2} \sum_{i_S \in \mathcal{I}_S} \delta(n(i_S)$$

Then if the data belongs to faces F_i of dimension k_i for the two arbitrary decomposable graphical models $J_i,\ i = 1, 2$ respectively, then, $d_{EDF} = k_1 - k_2$. We do not need facets for decomposable models. We just look at the cell counts.

A more sophisticated example

Let $V = \{a, b, c\}$, $\mathcal{D} = \{a, b, c, ab, bc\}$ and $I_a = \{0, 1, 2\} = I_b$ and $I_c = \{0, 1\}$. Thus I has $3 \times 3 \times 2 = 18$ elements and

$$J = \{100, 200, 010, 020, 001, 110, 210, 120, 220, 011, 021\}$$

has 11 elements with respective supports $a, a, b, b, c, ab, ab, ab, ab, bc, bc$. We have

$$\begin{aligned} \log p(201) &= \theta_0 + \theta_{200} + \theta_{001} \\ \log p(211) &= \theta_0 + \theta_{200} + \theta_{010} + \theta_{001} + \theta_{210} + \theta_{011} \\ \ldots &= \ldots \end{aligned}$$

and we can write

$$\log p = X\theta_J$$

where X has 18 rows and 11 columns.

The Zeta and Moebius function of $\lhd$

The order $j \lhd j'$ is a partial order on J.

It can be represented by its zeta function and its inverse can be represented by the Moebius function of the partial order.

$$\zeta(j, j') = 1 \text{ if } j \lhd j' \text{ and } 0 \text{ otherwise}$$

$$\mu(j, j') = (-1)^{|S(j')|-|S(j)|} \text{ if } j \lhd j' \text{ and } 0 \text{ otherwise}$$

Zeta function of the partial order on J_0

Let $J_0 = J \cup \{0\}$, the zeta function on J_0 can be read by columns on X_{J_0}

	000	100	200	010	020	110	210	120	220	001	011	021
000	1	0	0	0	0	0	0	0	0	0	0	0
100	1	1	0	0	0	0	0	0	0	0	0	0
200	1	0	1	0	0	0	0	0	0	0	0	0
010	1	0	0	1	0	0	0	0	0	0	0	0
020	1	0	0	0	1	0	0	0	0	0	0	0
110	1	1	0	1	0	1	0	0	0	0	0	0
210	1	0	1	1	0	0	1	0	0	0	0	0
120	1	1	0	0	1	0	0	1	0	0	0	0
220	1	0	1	0	1	0	0	0	1	0	0	0
001	1	0	0	0	0	0	0	0	0	1	0	0
011	1	0	0	1	0	0	0	0	0	1	1	0
021	1	0	0	0	1	0	0	0	0	0	0	1

Moebius function of the partial order on J_0

The columns of the matrix $X_{J_0}^{-1}$ gives $(X_{J_0}^{-1})_{j'j} = \mu(j, j')$

	000	100	200	010	020	110	210	120	220	001	011	021
000	1	0	0	0	0	0	0	0	0	0	0	0
100	-1	1	0	0	0	0	0	0	0	0	0	0
200	-1	0	1	0	0	0	0	0	0	0	0	0
010	-1	0	0	1	0	0	0	0	0	0	0	0
020	-1	0	0	0	1	0	0	0	0	0	0	0
110	1	-1	0	-1	0	1	0	0	0	0	0	0
210	1	0	-1	-1	0	0	1	0	0	0	0	0
120	1	-1	0	0	-1	0	0	1	0	0	0	0
220	1	0	-1	0	-1	0	0	0	1	0	0	0
001	-1	0	0	0	0	0	0	0	0	1	0	0
011	1	0	0	-1	0	0	0	0	0	-1	1	0
021	1	0	0	0	-1	0	0	0	0	-1	0	1

The facets can be read off $X_{J_0}^{-1}$

Indeed, $g_{j_0,D}$ can be read on the column j_0 and the rows j such that

$$S(j) = D \text{ and } j_0 \lhd j.$$

For example

$$\begin{aligned} g_{(200),ab} &= m_{200} - m_{210} - m_{220} \\ g_{0,ab} &= 1 - m_{100} - m_{200} - m_{010} - m_{020} \\ &\quad + m_{110} + m_{120} + m_{210} + m_{220} \end{aligned}$$

Some open problems

- Given the above framework, is a model decomposable if and only if the facets of C are the facets of the triangular type? Seth has a solution.
- If a graph is decomposable into prime components, are the facets of the graph the union of the facets of the prime components?
- How does one compute $I(m, \alpha)$ efficiently other than with the Laplace approximation?
- $\alpha^{-n} J_C(m)$ could be used as an approximation to $I(m, \alpha)$ for α small. How do you compute $J_C(m)$? Can we hope to find an analytic formula for $J_C(m)$ as in the decomposable case? Gerard is working on this.
- For regular models, when the data is on a face of C of dimension less than n, the BIC has to be modified and n is replaced by the dimension of the face. Matt Friedlander (Ph.D. student at York University) is working on the modification of BIC using the traditional methods of exponential families. Could he do it using the RLCT, avoiding to identify the facets of C?

Computation of $I(m, \alpha)$

$$I(m,\alpha) = \int_{R^n} \frac{\prod_{j\in J}(e^{\theta_j})^{\alpha m_j}}{\left(\sum_{i\in\mathcal{I}}\prod_{j\in J:j\lhd i}(e^{\theta_{ij}})^{x_{ij}}\right)^{\alpha}} d\theta$$

where $x_{ij}, i \in \mathcal{I}, j \in J$ are the entries of the $\mathcal{I} \times J$ submatrix of X with rows $f_i, i \in \mathcal{I}$.

Make the change of variable $v_j = \frac{e^{\theta_j}}{1+e^{\theta_j}}$ that is $e^{\theta_j} = \frac{v_j}{1-v_j}$.
Then

$$\sum_{i\in\mathcal{I}}\prod_{j\in J:j\lhd i}(e^{\theta_j})^{x_{ij}} = \frac{1}{\prod_{j\in J}(1-v_j)}\sum_{i\in\mathcal{I}}\prod_{l\in J} v_j^{x_{ij}}(1-v_j)^{(1-x_{ij})}$$

and

$$I(m,\alpha) = \int_{[0,1]^{|J|}} (\sum_{i\in\mathcal{I}}\prod_{l\in J} v_j^{x_{ij}}(1-v_j)^{(1-x_{ij})})^{-\alpha} \prod_{j\in J} v_j^{\alpha m_j - 1}(1-v_j)^{\alpha(1-m_j)-1} dv_j.$$

A numerical example

We will now illustrate our theorems using a well-known and much analysed data set. The data comes from a cross-classification of eight binary variables relating women's economic activity and husband's unemployment from a survey of households in Rochdale.

This study was conducted to elicit information about factors affecting the pattern of economic life and their time dynamics (Whittaker1990, p. 279). The variables are as follows:a, wife economically active (no,yes); b, age of wife > 38 (no,yes); c, husband unemployed (no,yes); d, child ≤ 4 (no,yes);e, wife's education, high-school+ (no,yes); f, husband's education, high-school+ (no,yes); g, Asian origin (no,yes); h, other household member working (no,yes).

There are 665 individuals cross-classified in 256 cells. The resulting table is sparse having 165 counts of zero.

Example: the table

To avoid the computation of high-dimensional normalizing constants in our example, we take as our data the four-dimensional $\{a, b, d, h\}$ marginal table given in the 4-dimensional table below, obtained from the Rochdale data.

The cells counts are written in lexicographical order with h varying fastest and a varying slowest, i.e. (0000,0001,0010,0011,. . .)

32	3	86	2	56	35	7	0
130	12	59	5	142	91	5	0

Example: The models considered

We will consider three models J_0, J_1 and J_2 such that

- J_0 is decomposable with cliques $\{a, d\}, \{d, b\}, \{b, h\}$ so that $\mathcal{D}_0 = \{a, b, d, h, (ad), (db), (bh)\}$. Since the data is binary, the dimension of this model is $|J_0| = 7$. The facets of C_0 are

$$1 - t_a - t_d + t_{ad} = 0 \quad t_a - t_{ad} = 0 \quad t_d - t_{ad} = 0 \quad t_{ad} = 0$$
$$1 - t_b - t_d + t_{bd} = 0 \quad t_b - t_{bd} = 0 \quad t_d - t_{bd} = 0 \quad t_{bd} = 0$$
$$1 - t_b - t_h + t_{bh} = 0 \quad t_b - t_{bh} = 0 \quad t_h - t_{bh} = 0 \quad t_{bh} = 0$$

Example: The models considered

- J_1 is a hierarchical model with generating set $\{(ad), (bd), (bh), (dh)\}$. This is not a graphical model and $\mathcal{D}_1 = \{a, b, d, h, (ad), (db), (bh), (dh)\}$. The dimension is $|J_1| = 8$. The facets of C_1 are given by

$$1 - t_a - t_d + t_{ad} = 0 \qquad t_a - t_{ad} = 0 \qquad t_d - t_{ad} = 0 \qquad t_{ad} = 0$$

and

$$t_h - t_{bh} - t_{dh} + t_{db} = 0 \quad 1 - t_b - t_d + t_{bd} = 0 \quad t_b - t_{bd} = 0 \quad t_d - t_{bd} = 0 \quad t_{bd} = 0$$
$$t_d - t_{dh} - t_{bh} + t_{bh} = 0 \quad 1 - t_b - t_h + t_{bh} = 0 \quad t_b - t_{bh} = 0 \quad t_h - t_{bh} = 0 \quad t_{bh} = 0$$
$$t_b - t_{bh} - t_{bd} + t_{dh} = 0 \quad 1 - t_d - t_h + t_{dh} = 0 \quad t_d - t_{dh} = 0 \quad t_h - t_{dh} = 0 \quad t_{dh} = 0$$

Example: The models considered

- J_2 is decomposable with cliques $\{b, d, h\}, \{a\}$ and $\mathcal{D}_2 = \{a, b, d, h, (ad), (db), (bh), (dh), (bdh)\}$. The dimension is $|J_2| = 8$. The facets of C_2 are given by

$$1 - t_a - t_d + t_{ad} = 0 \qquad t_a - t_{ad} = 0 \qquad t_d - t_{ad} = 0 \qquad t_{ad} = 0$$

and

$$\begin{aligned}
1 - t_b - t_d + t_{bd} &= 0 & t_b - t_{bd} &= 0 & t_d - t_{bd} &= 0 & t_{bd} &= 0 \\
1 - t_b - t_h + t_{bh} &= 0 & t_b - t_{bh} &= 0 & t_h - t_{bh} &= 0 & t_{bh} &= 0 \\
1 - t_d - t_h + t_{dh} &= 0 & t_d - t_{dh} &= 0 & t_h - t_{dh} &= 0 & t_{dh} &= 0 \\
t_{bdh} &= 0 & & & & & & \qquad (2)
\end{aligned}$$

Example: The marginal counts

Since the data is binary, we can identify the set J to the set $\mathcal{D}$ and we will therefore denote the canonical statistic t as

$$t = (t(D),\ D \in \mathcal{D}).$$

From the table , we can obtain the marginal counts

$$t_a = 444,\ t_b = 336,\ t_d = 164,\ t_h = 148,\ t_{bd} = 12,\ t_{bh} = 126,\ t_{dh} = 7,\ t_{bdh} = 0.$$

We therefore see that since $t_{bdh} = 0$, the data belongs to one facet of J_2 and no facet of either C_0 or C_1. F

$$\begin{aligned} B_{1,0} &\sim \alpha^{k_1 - k_0} = \alpha^{8-7} = \alpha^1 \\ B_{2,0} &\sim \alpha^{k_2 - k_0} = \alpha^{8-8} = 1 \end{aligned}$$

Example: Plots

The numerical computations of $B_{1,0}$ and $B_{2,0}$ illustrate the different behaviours of these two Bayes factors.

Figure 1: Convergence plot of model parameters