Geometry of hierarchical discrete loglinear models for Bayes factors

Hélène Massam

York University

with G. Letac, Université Paul Sabatier

The problem

- The data is given by a $|V|$-dimensional contingency table classifying N individuals according to V criteria.
- We consider the class of hierarchical loglinear models.
- The cell counts follow a multinomial distribution with density $f(t;\theta) = e^{\langle \theta, t\rangle - Nk(\theta)}$.
- The conjugate prior for θ is of the form $\pi(\theta) = \frac{e^{\alpha\langle \theta, m\rangle - \alpha k(\theta)}}{I(m,\alpha)}$.
- The Bayes factor between model 1 and model 2 is

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

- We study the behaviour of $B_{1,2}$ as $\alpha \to 0$.

Objects of interest

- the generating measure μ for the multinomial distribution
- the convex hull C of the support of μ
- The characteristic function $\mathbb{J}_C$ of the convex polytope C
- The polar set of C
- the face of $\overline{C}$ containing the data and its dimension k.

<u>The result</u>

$$B_{1,2} \sim \alpha^{k_1 - k_2}.$$

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

- 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

$$\log p(i) = \theta_\emptyset + \sum_{j \in J_i} \theta_j \quad \text{with} \quad \log p(0) = \theta_0.$$

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
$L(\theta_J) = (1 + \sum_{i \in I \setminus \{0\}} \exp \sum_{j \in J_i} \theta_j)$.

The measure generating the multinomial

Let $(e_j, j \in J)$ be the canonical basis of R^J and let $f_i = \sum_{j\in J, j \triangleleft i} e_j, \quad i \in I.$ For $G = a - - - - b - - - - c$

$\mathcal{D}$	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 $\theta \in R^J$,

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

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:
$f_i, i \in I$ 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
- $\alpha(m_j, \ \ j \in J)$ 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 $\lim \frac{I(m_2,\alpha)}{I(m_1,\alpha)}$.

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

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

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

- $J=\{(1,0,0),(0,1,0),(0,0,1),(1,1,0),(0,1,1)\}$ with C spanned by $f_j, j\in J$ and $m=\sum_{j\in J} m_j f_j$. Then

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

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

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

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

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

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 $\mathrm{vol}_{n-1}\widehat{F}_\epsilon \sim c\epsilon^k$
- $$\int_{\overline{C^o}} \frac{d\theta}{(1-(1-\lambda)\langle\theta, y\rangle)^{n+1}} = \int_0^\infty \frac{\mathrm{vol}_{n-1}\widehat{F}_\epsilon d\epsilon}{(1-(1-\lambda)(1-\epsilon))^{n+1}} = \int_0^\infty \frac{f(\epsilon)d\epsilon}{(1-(1-\lambda)(1-\epsilon))^{n+1}}$$
- 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

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

and the affine forms

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

Some facets of C

All subsets of the form

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

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

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

<u>Example</u> $a - - - b - - - c$. 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 faces of C.

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

$$\begin{aligned} t_{ab} = 0, j = (1,1,0); &\quad t_a - t_{ab} = 0, j = (1,0,0) \\ t_b - t_{ab} = 0, j = (0,1,0); &\quad 1 - t_a - t_b + t_{ab} = 0, \; S(j) = \emptyset \\ t_{bc} = 0, j = (0,1,1); &\quad t_b - t_{bc} = 0, j = (0,1,0) \\ t_c - t_{bc} = 0, j = (0,0,1); &\quad 1 - t_b - t_c + t_{bc} = 0, \; S(j) = \emptyset. \end{aligned}$$

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

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

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.