Bayes regularization and the geometry of discrete hierarchical loglinear models

Hélène Massam

York University

with G. Letac, Université Paul Sabatier

The problem

- We want to fit a hierarchical loglinear model to some discrete data given under the form of a contingency table.
- We put the Diaconis-Ylvisaker conjugate prior on the loglinear parameters of the multinomial distribution for the cell counts of the contingency table.
- We study the behaviour of the Bayes factor as the hyperparameter α of the conjugate prior tends to 0
- We are led to study the convex hull C of the support of the multinomial distribution.
- The facets of C are the most important objects in this study.

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 loglinear model

- We choose a special cell $0 = (0, \ldots, 0)$.
- The set $\mathcal{D} = \{D \subseteq V : \ D_1 \subset D \Rightarrow D_1 \in \mathcal{D}\}$ define the hierarchical loglinear model.

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

- We define $S(i) = \{\gamma \in V : \ i_\gamma \neq 0\}$ and

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

- We change parametrization

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

The loglinear model:cont'd

- 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

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

Example

Consider the hierarchical model with

$$V = \{a, b, c\},\ \mathcal{A} = \{\{a, b\}, \{b, c\}\},\ I_a = \{0, 1, 2\} = I_b,\ \ I_c = \{0, 1\},$$

and $i = (0, 2, 1)$. We have

$$\begin{aligned}
\mathcal{D} &= \{a, b, c, ab, bc\} \\
J &= \{(1,0,0),(2,0,0),(0,1,0),(0,2,0),(0,0,1),(1,1,0),(1,2,0), \\
&\qquad (2,1,0),(2,2,0),(0,1,1),(0,2,1)\} \\
J_i &= \{(0,2,0),(0,0,1),(0,2,1)\} \\
\log p(0,2,1) &= \theta^{\emptyset}_{(0,2,1)} + \theta^{b}_{(0,2,1)} + \theta^{c}_{(0,2,1)} + \theta^{b,c}_{(0,2,1)} \\
&= \theta_{(0,0,0)} + \theta_{(0,2,0)} + \theta_{(0,0,1)} + \theta_{(0,2,1)} \\
&= \theta_0 + \sum_{j \in J_i} \theta_j
\end{aligned}$$

The multinomial hierarchical model

Since $J = \cup_{i \in \mathcal{I}} J_i$, the loglinear parameter is

$$\theta_J = (\theta_j, \;\; j \in J).$$

The hierarchical model is characterized by J. For $i \neq 0$, the loglinear model can then be written

$$\log p(i) = \theta_0 + \sum_{j \in J_i} \theta_j$$

with $\log p(0) = \theta_0$. Therefore

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

and

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

The model as an exponential family

Make the change of variable

$$(n) = (n(i), i \in I\backslash\{0\}) \mapsto t = (t(i_E) = n(i_E), E \subseteq V\backslash\{\emptyset\}, i \in I\backslash\{0\}).$$

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

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

and $L(\theta_J) = (1 + \sum_{i \in I\backslash\{0\}} \exp\sum_{j \in J_i} \theta_j)$.
It is an NEF of dimension $|J|$, generated by the following measure.

The generating vectors

The set of functions from J to R is denoted by R^J and we write any function $h \in R^J$ as $h = (h(j), j \in J)$, which we can think of as a $|J|$ dimensional vector in $R^{|J|}$. Let $(e_j, j \in J)$ be the canonical basis of R^J and let

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

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

The measure

We note that in our example R^I is of dimension 8 while R^J is of dimension 5 and the $(f_j, j \in J)$ are, of course, 5-dimensional vectors. Consider now the counting measure in R^J

$$\mu_J = \delta_0 + \sum_{i \in \mathcal{I}} \delta_{f_i}.$$

For $\theta \in R^J$, the Laplace transform of μ_J is

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

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: The convex hull of the support of μ_J

Since $\mu_J = \delta_0 + \sum_{i \in \mathcal{I}} \delta_{f_i}$,

C_J is the open convex hull of $0 \in R^J$ and $f_i, i \in I$.

It is important to identify this convex hull since Diaconis and Ylvisaker (1974) have proven that the conjugate prior to an NEF, defined by

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

is proper when the hyperparameters $m_J \in R^J$ and $\alpha \in R$ are such that

$$\alpha > 0 \text{ and } m_J \in C_J.$$

The DY conjugate prior

Clearly, we can write the multinomial density as $f(t_J|\theta_J) = f(t_J|\theta_J, J)$ where J represents the model. Assuming we put a uniform discrete distribution on the set of models,the joint distribution of J, t_J, θ_J is

$$f(J, t_J, \theta_J) \propto \frac{1}{I(m_J, \alpha)} e^{\{\langle \theta_J, t_J + \alpha m_J \rangle - (\alpha + N) \log L(\theta_J)\}}$$

and therefore the posterior density of J given t_J is

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

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

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

$$\frac{I(m_2, \alpha)}{I(m_1, \alpha)} \times \frac{I\left(\frac{t_1+\alpha m_1}{\alpha+N}, \alpha+N\right)}{I\left(\frac{t_2+\alpha m_2}{\alpha+N}, \alpha+N\right)}.$$

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 $\lim \frac{I(m_2,\alpha)}{I(m_1,\alpha)}$ when $\alpha \to 0$.

When $\alpha \to 0$, if $\frac{t_k}{N} \in \bar{C}_k \setminus C_k,\ k = 1, 2$, then, we have to worry about both limits.

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

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

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

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

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

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

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

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

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

Limit of the Bayes factor

Let models J_1 and J_2 be such that $|J_1| > |J_2|$ and the marginal counts $\frac{t_i}{N}$ are both in C_i. 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.

Some comments

If $\frac{t_i}{N}$ are both in $C_i, i = 1, 2$ and $|J_1| \neq |J_2|$, we need not compute $J_C(m)$.

If $\frac{t_i}{N}$ are both in $C_i, i = 1, 2$ and $|J_1| = |J_2|$, then we might want to compute $J_C(m_i) i = 1, 2$. In this case, we have a few theoretical results. We define the polar convex set C_0 of C

$$C^0 = \{\theta \in R^n \; ; \; \langle \theta, x \rangle \leq 1 \;\; \forall x \in C\}$$

then

- $\frac{J_C(m)}{n!} =$ Vol$(C - m)^0$
- 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 $< K$.

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

We now consider the case when $\frac{t}{N}$ belongs to the boundary of C. Then each facet of $\bar{C}$ (of dimension $|J| - 1$) is of the form

$$F_g = \{x \in \bar{C} : g(x) = 0\}$$

where g be an affine form on R^J.

Theorem

Suppose $\frac{t}{N} \in \bar{C} \setminus C$ belongs to exactly M faces of $\bar{C}$. Then

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

exists and is positive.

The Bayes factor

Combining the study of the asymptotic behaviour of $I(m, \alpha)$ and $I(\frac{\alpha m+t}{\alpha+N}, \alpha + N)$, we obtain that

when $\alpha \to 0$, the Bayes factor behaves as follows

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

$$\sim \quad C\alpha^{|J_1|-|J_2|-[\min(M_1,|J_1|)-\min(M_2,|J_2|)]} \frac{J_{C_1}(m_1)}{J_{C_2}(m_2)}$$

where C is a positive constant.

Facets of C when G is decomposable

Let $(C_i, i = 1, \dots, k)$ and $(S_i, i = 2, \dots, k)$ be the set of cliques and separators of G. Then

$$I(m,\alpha) = \frac{\prod_{C\in\mathcal{C}} \Gamma(\alpha g_{0,C}(m)) \prod_{\{j\in J; S(j)\subset C\}} \Gamma(\alpha g_{j,C}(m))}{\Gamma(\alpha) \prod_{S\in\mathcal{S}} \left[\Gamma(\alpha g_{0,S}(m)) \prod_{\{j\in J; S(j)\subset S\}} \Gamma(\alpha g_{j,S}(m))\right]^{\nu(S)}}.$$

where for $D \in \mathcal{C}$ or $D \in \mathcal{S}$,

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

Facets of C when G is decomposable

Since $\Gamma(z) \sim \frac{1}{z}$ when $z \longrightarrow 0$, we have

$$I(m,\alpha)^{-1} \quad \sim \quad \alpha^{|J|} \frac{\prod_{C\in\mathcal{C}} g_{0,C}(m) \prod_{\{j\in J; S(j)\subset C\}} g_{j,C}(m)}{\prod_{S\in\mathcal{S}} \left[g_{0,S}(m) \prod_{\{j\in J; S(j)\subset S\}} g_{j,S}(m)\right]^{\nu(S)}}$$

i.e.

$$\alpha^{|J|} I(m,\alpha) \quad \rightarrow \quad \frac{\prod_{S\in\mathcal{S}} \left[g_{0,S}(m) \prod_{\{j\in J; S(j)\subset S\}} g_{j,S}(m)\right]^{\nu(S)}}{\prod_{C\in\mathcal{C}} g_{0,C}(m) \prod_{\{j\in J; S(j)\subset C\}} g_{j,C}(m)}$$

Therefore the facets of C are the intersection of C with the hyperplanes

$$g_{0,C}(m) = 0, \quad g_{j,C}(m),\ j \in J; S(j) \subset C,\ C \in \mathcal{C}.$$

Example

For G as $a - - - b - - - c$ and for binary data,
the faces of C are

$$m_{ab} = 0,\ m_a - m_{ab} = 0,\ m_b - m_{ab} = 0,\ 1 - m_a - m_b + m_{ab} = 0$$

and

$$m_{bc} = 0,\ m_b - m_{bc} = 0,\ m_c - m_{bc} = 0,\ 1 - m_b - m_c + m_{bc} = 0.$$

Facets common to all C's

Let $\mathcal{C}$ be the set of generators of the hierarchical model.

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

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

The intersection of C with the hyperplanes

$$g_{0,D}(m) = 0, \quad g_{j_0,D}(m),\ S(j_0) \subset D,\ D \in \mathcal{C}$$

are facets of C whatever the hierarchical model. This is a new result since it gives us, for example, the facets of the hierarchical model with 4 factors a, b, c, d and three-way interaction $(abc), (bcd), (cda), (dab)$.

Facets of C when G is a cycle

Let V be the set of vertices of G and let E be the set of cliques. For any subset $F \subseteq E$ with odd cardinality $|F|$,

$$\sum_{(a,b)\in F} (m_a + m_b - 2m_{ab}) - \Big(\sum_{v\in V} m_v - \sum_{e\in E} m_e \Big) \leq \frac{|F|-1}{2} \,. \quad (1)$$

This result can be deduced from known results in geometry on the facets of a correlation polytope governed by a graph.

Example

Consider the hierarchical model with $\mathcal{D} = \{(ab), (bc), (ca)$.

The 16 facets are given by the following affine forms being equal to 0:

$$
\begin{array}{ccc}
m_{ab} & m_{bc} & m_{ac} \\
m_a - m_{ab} & m_b - m_{bc} & m_c - m_{ac} \\
m_b - m_{ab} & m_c - m_{bc} & m_a - m_{ac} \\
1 - m_a - m_b + m_{ab} & 1 - m_b - m_c + m_{bc} & 1 - m_a - m_c + m_{ac} \\
m_c - m_{ac} - m_{bc} + m_{ab} & m_a - m_{ab} - m_{ac} + m_{bc} & m_b - m_{ab} - m_{bc} + m_{ac} \\
 & 1 - m_a - m_b - m_c + m_{ac} + m_{ab} + m_{bc} &
\end{array}
$$

Bayesian networks

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 $\mathcal{A} = \{(a, b, c)\}$.
$|J_2| = 7$

Our results

Model $\mathcal{M}_2$: $a - - - - b - - - - c$: $|J_2| = 5$. The faces expressed in traditional notation are

$$n_{11+} = n_{10+} = n_{01+} = n_{00+} = n_{+11} = n_{+10} = n_{+01} = n_{+00} = 0$$

Model $\mathcal{M}_1$: $|J_1| = 7$. The faces expressed in traditional notation are

$$n_{000} = n_{100} = n_{010} = n_{001} = n_{110} = n_{011} = n_{101} = n_{111} = 0$$

Example The data is such that $n_{000} = n_{100} = n_{101} = 0$. Therefore in $\mathcal{M}_1$, $\frac{t_1}{N}$ belongs to $M_1 = 3$ faces and in $\mathcal{M}_2$, $\frac{t_2}{N}$ belongs to $M_1 = 2$ faces $n_{10+} = 0 = n_{+00}$.
Thus the Bayes factor $\sim \alpha^d$ where

$$d = |J_1| - |J_2| - [min(|J_1, M_1) - min(|J_2|, M_2)] = 7 - 5 - [3 - 2] = 1$$

Steck and Jaakola (2002)

Define the effective degrees of freedom to be

$$d_{EDF} = \sum_{i} I(n_i) - \sum_{i_{ab}} I(n(i_{ab})) - \sum_{i_{bc}} I(n(i_{bc})) + \sum_{i_b} I(n(i_b))$$

<u>Theorem</u> If $d_{EDF} > 0$, the Bayes factor tends to 0 and if $d_{EDF} < 0$ the Bayes factor tends to $+\infty$. If $d_{EDF} = 0$, the Bayes factor can converge to any value.

In our example

$$d_{EDF} = 5 - 3 - 3 + 2 = 1$$

Our results agree with SJ in the particular case of Bayesian networks. Our results give a much finer analysis for a more general class of problems.

Example of model search

We study the Czech Autoworkers 6-way table from Edwards and Havranek (1985).

This cross-classfication of 1841 men considers six potential risk factors for coronary trombosis:

- a, smoking;
- b, strenuous mental work;
- c, strenuous physical work;
- d, systolic blood pressure;
- e, ratio of beta and alpha lipoproteins;
- f, family anamnesis of coronary heart disease.

Edwards and Havranek (1985) use the LR test and Dellaportas and Forster (1999) use a Bayesian search with normal priors on the θ to analyse this data.

Czech Autoworkers example our method

We use a Bayesian search with

- MC^3
- our prior with $\alpha = 1, 2, 3, 32$ and then $\alpha = .05, .01$ and equal fictive counts for each cell
- The Laplace approximation to the marginal likelihood

Czech Autoworkers example

Search	$\alpha = 1$		$\alpha = 2$	
Dec.	$bc\|ace\|ade\|f$	0.250	$bc\|ace\|ade\|f$	0.261
	$bc\|ace\|de\|f$	0.104	$bc\|ace\|de\|f$	0.177
	$bc\|ad\|ace\|f$	0.102	$bc\|ace\|de\|bf$	0.096
	$ac\|bc\|be\|de\|f$	0.060	$bc\|ad\|ace\|f$	0.072
	$bc\|ace\|de\|bf$	0.051	$bc\|ace\|de\|bf$	0.065
	$bc\|ace\|de\|f$	med	$bc\|ad\|ace\|de\|f$	med
Graph.	$ac\|bc\|be\|ade\|f$	0.301	$ac\|bc\|be\|ade\|f$	0.341
	$ac\|bc\|ae\|be\|de\|f$	0.203	$ac\|bc\|be\|ade\|bf$	0.141
	$ac\|bc\|be\|ade\|bf$	0.087	$ac\|bc\|ae\|be\|de\|f$	0.116
	$ac\|bc\|ad\|ae\|be\|f$	0.083	$ac\|bc\|be\|ade\|ef$	0.059
	$ac\|bc\|ae\|be\|de\|bf$	0.059		
	$ac\|bc\|ad\|ae\|be\|de\|f$	med	$ac\|bc\|be\|ade\|f$	med
Hierar.	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.241	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.175
	$ac\|bc\|ad\|ae\|be\|de\|f$	0.151	$ac\|bc\|ad\|ae\|be\|de\|f$	0.110
	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.076	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.078
	$ac\|bc\|ad\|ae\|ce\|de\|bf$	0.070	$ac\|bc\|ad\|ae\|ce\|de\|bf$	0.072
	$ac\|bc\|ad\|ae\|ce\|de\|f$	med	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	med

Results for α close to 0

Search	$\alpha = .5$		$\alpha = .01$	
Hierar.	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.3079	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.2524
	$ac\|bc\|ad\|ae\|be\|de\|f$	0.1926	$ac\|bc\|ad\|ae\|be\|de\|f$	0.1577
	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.0686	$ac\|bc\|ae\|ce\|de\|f$	0.1366
	$ac\|bc\|ad\|ae\|ce\|de\|be$	0.0631	$ac\|bc\|d\|ae\|ce\|f$	0.1168
	$ac\|bc\|ad\|ae\|ce\|de\|f$	med	$ac\|bc\|ae\|de\|f$	0.0854
			$ac\|bc\|c\|ae\|be\|f$	0.0730
			$ac\|bc\|ad\|ae\|ce\|f$	0.0558

Recall that for $\alpha = 1, 2$, the most probable model was $ac|bc|ad|ae|ce|de|f$ with respective probablities 0.241 and 0.175.

As $\alpha \mapsto 0$, the models become sparser but are consistent with those corresponding to larger values of α.

Another example

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

Marginal a, b, d, h table from the Rochdale data in whittaker1990. The cells counts are written in lexicographical order with h varying fastest and a varying slowest.

The three models considered

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

(a) J_0 is decomposable with cliques $\{a,d\},\{d,b\},\{b,h\}$ so that $\mathcal{D}$ as defined in Section 2 is

$$\mathcal{D}_0 = \{a,b,d,h,(ad),(db),(bh)\},\ |J_0| = 7,\ M_0 = 0.$$

(b) J_1 is a hierarchical model with generating set $\{(ad),(bd),(bh),(dh)\}$. This is not a graphical model and

$$\mathcal{D}_1 = \{a,b,d,h,(ad),(db),(bh),(dh)\},\ |J_1| = 8\ M_1 = 0.$$

(c) J_2 is decomposable with cliques $\{b,d,h\},\{a\}$,and

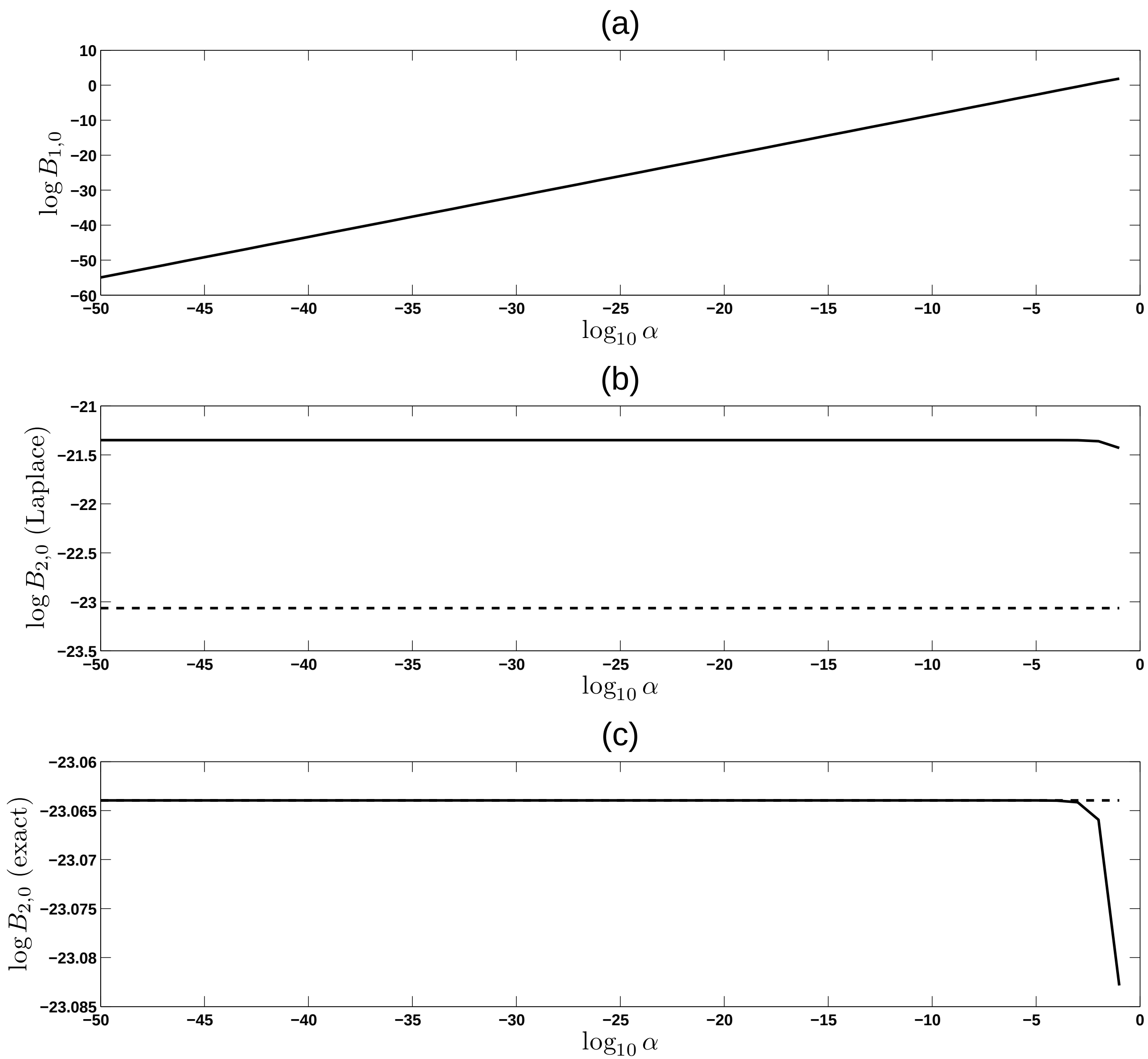
$$\mathcal{D}_2 = \{a,b,d,h,(ad),(db),(bh),(dh),(bdh)\},\ |J_2| = 8,\ M_2 = 1.$$

Asymptotics of $B_{1,0}$ and $B_{2,0}$

We have

$$\begin{aligned}
B_{1,0} &\sim \alpha^{|J_0|-|J_1|-[\min(M_0,|J_0|)-\min(M_1,|J_1|)]}\frac{J_{C_1}(m_1)}{J_{C_0}(m_0)} \\
&= C_{1,0}\alpha^{(7-8-(0-0)} = C\alpha^{-1} \\
B_{2,0} &\sim \alpha^{|J_0|-|J_2|-[\min(M_0,|J_0|)-\min(M_2,|J_2|)]}\frac{J_{C_2}(m_2)}{J_{C_0}(m_0)} \\
&= C_{2,0}\alpha^{(7-8-(0-1)} = C_{2,0}\alpha^{0} = C_{2,0}
\end{aligned}$$

The graphs



(a)
10
0
−10
−20
−30
−40
−50
−60
$\log B_{1,0}$
−50
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
$\log_{10} \alpha$
(b)
−21
−21.5
−22
−22.5
−23
−23.5
$\log B_{2,0}$ (Laplace)
−50
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
$\log_{10} \alpha$
(c)
−23.06
−23.065
−23.07
−23.075
−23.08
−23.085
$\log B_{2,0}$ (exact)
−50
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
$\log_{10} \alpha$