The Annals of Statistics
0, Vol. 0, No. 00, 1–30
DOI: 10.1214/12-AOS974

BAYES FACTORS AND THE GEOMETRY OF DISCRETE HIERARCHICAL LOGLINEAR MODELS

BY GÉRARD LETAC AND HÉLÈNE MASSAM[1]

Université Paul Sabatier and York University

A standard tool for model selection in a Bayesian framework is the Bayes factor which compares the marginal likelihood of the data under two given different models. In this paper, we consider the class of hierarchical loglinear models for discrete data given under the form of a contingency table with multinomial sampling. We assume that the prior distribution on the loglinear parameters is the Diaconis–Ylvisaker conjugate prior, and the uniform is the prior distribution on the space of models. Under these conditions, the Bayes factor between two models is a function of the normalizing constants of the prior and posterior distribution of the loglinear parameters. These constants are functions of the hyperparameters (m, α) which can be interpreted, respectively, as the marginal counts and total count of a fictive contingency table.

We study the behavior of the Bayes factor when α tends to zero. In this study, the most important tool is the characteristic function $\mathbb{J}_C$ of the interior C of the convex hull $\overline{C}$ of the support of the multinomial distribution for a given hierarchical loglinear model. If h_C is the support function of C, the function $\mathbb{J}_C$ is the Laplace transform of $\exp(-h_C)$. We show that, when α tends to 0, if the data lies on a face F_i of $\overline{C}_i, i = 1, 2$, of dimension k_i, the Bayes factor behaves like $\alpha^{k_1-k_2}$. This implies in particular that when the data is in C_1 and in C_2, that is, when k_i equals the dimension of model J_i, the sparser model is favored, thus confirming the idea of Bayesian regularization.

In order to find the faces of $\overline{C}$, we need to know its facets. We show that since here C is a polytope, the denominator of the rational function $\mathbb{J}_C$ is the product of the equations of the facets. We also identify a category of facets common to all hierarchical models for discrete variables, not necessarily binary. Finally, we show that these facets are the only facets of $\overline{C}$ when the model is graphical with respect to a decomposable graph.

1. Introduction.

We consider data given under the form of a contingency table representing the classification of N individuals according to a finite set of criteria. We assume that the cell counts in the contingency table follow a multinomial distribution. We also assume that the cell probabilities are modeled according to a hierarchical loglinear model (henceforth called hierarchical model). The multinomial distribution for the hierarchical model is a natural exponential family of

Received March 2011; revised January 2012.

[1]Supported by NSERC Discovery Grant A8947.

MSC2010 subject classifications. 62H17, 62C10, 62J12.

Key words and phrases. Bayes, sparse contingency tables, conjugate priors, characteristic function of a convex set, Dirichlet distribution.

the general form $L(\theta)^{-1}\exp\langle\theta, t\rangle\mu(dt)$ where μ is the generating measure and L is its Laplace transform. The Diaconis–Ylvisaker [7] (henceforth abbreviated DY) conjugate prior is the probability

$$I(m,\alpha)^{-1}L(\theta)^{-\alpha}\exp(\alpha\langle\theta, m\rangle)\,d\theta, \tag{1.1}$$

where m and α are hyperparameters and $I(m,\alpha)$ is the normalization constant. Massam et al. [16] have identified and studied the Diaconis–Ylvisaker conjugate prior for the so called baseline constrained loglinear parametrization of the multinomial for hierarchical models. This prior is a generalization of the hyper Dirichlet defined by Dawid and Lauritzen [5] for graphical models Markov with respect to decomposable graphs. Since decomposable graphical models, and more generally graphical models, form a subclass of the class of hierarchical models we will call this prior the generalized hyper Dirichlet. For the generalized hyper Dirichlet or the hyper Dirichlet, α is a positive scalar while m is a vector. The scalar α can be interpreted as the total sample size of a fictive contingency table, and m can be interpreted as the vector of various marginal counts of the same table. It is therefore traditional to take α small relatively to the total data count N. In this paper, we will use the loglinear parametrization for the hierarchical model and the generalized hyper Dirichlet as the prior, as defined in [16].

In a Bayesian framework, the Bayes factor is one of the main tools for model selection in the class of hierarchical models. The aim of this paper is to study the behavior of the Bayes factor for the comparison of two hierarchical models J_1 and J_2 when α is very small, that is, when $\alpha \to 0$. The motivation for this study is twofold. First, it has been observed that as $\alpha \to 0$, in general, the Bayes factor will select the sparser model, that is, the model with the parameter space of smallest dimension or equivalently the model with the least number of interactions. This is commonly called the phenomenon of regularization. Second, Steck and Jaakkola ([20], Proposition 1) have shown that, however, this is not always the case and that, in fact, the behavior of the Bayes factor between two Bayesian networks differing by one edge only depends upon a quantity which they call d_{EDF}, effective degrees of freedom, and which depends solely on the data. Comparing two such Bayesian networks is equivalent to comparing two graphical models on three variables, the saturated model and the model Markov with respect to the two-link chain, with one conditional independence. It is therefore natural to seek a generalization of the results in [20] when two arbitrary hierarchical models are considered.

Our aim is to formally explain when the sparser model is selected, when it is not and why. We also want to develop tools to predict what the behavior of the Bayes factor will be, for two given models.

Since in the case of the DY conjugate prior, the posterior probability of model J given the data is equal to the ratio of the posterior and prior normalizing constants, we will be led to study the asymptotic behavior, as $\alpha \to 0$, of the normalizing constant $I(m,\alpha)$ in (1.1). In this study, one important mathematical object will surface. The multinomial distribution for a given hierarchical model J is a natural exponential family. We denote by C the interior of the convex hull $\overline{C}$ of the support of the measure generating this multinomial distribution. The position of the data with respect to C, that is, whether the data is in C or on one of the faces of $\overline{C}$, will determine the behavior of the Bayes factor. The important object is the characteristic function $\mathbb{J}_C$ of this polytope C, defined in (3.1); $\mathbb{J}_C(m)$ is also defined in the literature as $n!$ times the volume of the polar set of $C - m$; see [3]. It is through $\mathbb{J}_C$ that we will be able to find the asymptotic behavior of $I(m, \alpha)$. Our central statistical result is that, as $\alpha \to 0$, the Bayes factor $B_{1,2}$ between two hierarchical models J_1 and J_2 behaves as follows:

$$B_{1,2} \sim D\alpha^{k_1 - k_2}, \tag{1.2}$$

where D is a positive constant and k_i, $i = 1, 2$, are, respectively, the dimension of the face of $\overline{C}_i$ containing the data in its relative interior. When the data is in both the open convex sets C_i, $i = 1, 2$, we have of course that

$$B_{1,2} \sim D\alpha^{|J_1| - |J_2|},$$

where $|J|$ denotes the dimension of the model, and this explains that in general the Bayes factor favors the sparser model since, in general for low-dimensional tables, the data is in the open polytope C_i. However with modern genetic or sociological data, we often deal with very sparse high-dimensional tables. In that case, the data may well be on a face of dimension $k_i < |J_i|$. Then, as shown in [20] for three-factor models, the sparser model is not necessarily favored by the Bayes factor. We do not consider, in this paper, the case $\alpha \to +\infty$ since in that case, the behavior of $I(m, \alpha)$ is well known; see, for example, [19] or [12].

The contents of the paper are as follows. In Section 2, we give the matrix representation of the hierarchical loglinear model that we are going to work with, and we recall the form of the multinomial and the DY conjugate prior for that model. In Section 3, we show that since C is a polytope, the function $\mathbb{J}_C$ is a quotient of polynomials and its denominator is the product of the equations of the facets of $\overline{C}$. We also give the basic theorems on the behavior of $I(m, \alpha)$ and $\mathbb{J}_C(m)$ when m goes close to the boundary of C. In Section 4, we give our main statistical results and relate them to those in [20]. In Section 5, we give a category of facets of C common to all hierarchical models. We also show that these are the only facets in the case of a decomposable graphical model.

Some of the proofs are given in the paper and some in the supplementary file [15]. For ease of reference, the numbering in the supplementary file [15] is exactly the same as in the paper.

2. Preliminaries.

2.1. *The hierarchical model.* While we keep the traditional notation as given in [5] for cells and cell counts of the contingency table, we simplify the notation introduced in [16] for the set of nonzero loglinear parameters.

Let V be a finite set of indices representing $|V|$ criteria. We assume that the criterion labeled by $v \in V$ can take values in a finite set I_v. We consider N individuals classified according to these $|V|$ criteria. The resulting counts are gathered in a contingency table such that

$$I = \prod_{v \in V} I_v$$

is the set of cells $i = (i_v, v \in V)$. If $D \subset V$ and $i \in I$, we write $i_D = (i_v, v \in D)$ for the D-marginal cell. We write $\mathbb{R}^I$ for the space of real functions $i \mapsto x(i)$ defined on I. The element $x \in \mathbb{R}^I$ is seen sometimes as a vector, sometimes as the function $i \mapsto x(i)$ on I.

Let $\mathcal{D}$ be a family of nonempty subsets of V such that $D \in \mathcal{D}$, $D_1 \subset D$ and $D_1 \neq \varnothing$ implies $D_1 \in \mathcal{D}$. In order to avoid trivialities we assume $\bigcup_{D \in \mathcal{D}} D = V$. In the literature such a family $\mathcal{D}$ is called a hypergraph (see [14]) or an abstract simplicial complex (see [9]) or more simply the generating class (see [8]). Following the notation introduced in [4], we denote by $\Omega_{\mathcal{D}}$ the linear subspace of $x \in \mathbb{R}^I$ such that there exist functions $\theta_D \in \mathbb{R}^I$ for $D \in \mathcal{D}$ depending only on i_D and such that $x = \sum_{D \in \mathcal{D}} \theta_D$, that is,

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

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 $\Omega_{\mathcal{D}}$

$$\log p(i) = \theta_{\varnothing} + \sum_{D \in \mathcal{D}} \theta_D(i_D), \tag{2.1}$$

where $\theta_{\varnothing}$ does not depend on i and is thus a constant.

Needless to say, representation (2.1) is not unique. In order to make it unique, we need to impose certain constraints on the parameters $\theta(i_D), i_D \in \mathcal{I}_D, D \in \mathcal{D}$. To this end, we first select a special element in each I_v. For convenience we denote it 0. By abuse of notation, we also denote 0 in I the cell with all its components equal to 0. This special element in I_v is denoted r_v in [4] and i^* in [14] and [16], but we find the notation 0 more convenient. Actually the choice of the special element 0 in each I_v is arbitrary and does not affect our results. It has been proved in [4] and later more explicitly in [14], Proposition B.4 and formula (B.11), that representation (2.1) holds and is unique if we impose the constraints that, for $D \in \mathcal{D}$,

$$\text{if } i_v = 0 \quad \text{for some } v \in D \quad \text{then } \theta_D(i_D) = 0. \tag{2.2}$$

Using (2.2), representation (2.1) becomes

$$\log p(i) = \theta_{\varnothing} + \sum_{D \in \mathcal{D}, i_v \neq 0, \forall v \in D} \theta_D(i_D). \tag{2.3}$$

To reach a more concise notation, we are led to define the support $S(i)$ of a cell i as

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

and the particular subset J of I as follows:

$$J = \{j \in I, S(j) \in \mathcal{D}\}. \tag{2.4}$$

We see immediately that for a given $D \in \mathcal{D}$ and for a given $\theta_D(i_D)$ such that $i_\gamma \neq 0, \forall \gamma \in D$, there is only one $j \in J$ such that $S(j) = D$ and $j_D = j_{S(j)} = i_D$ and conversely. We can therefore write

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

The unique representation (2.3) of $\log p \in \Omega_{\mathcal{D}}$ is therefore given by the *free* parameters

$$\{\theta_j, j \in J\}, \tag{2.5}$$

and (2.3) becomes

$$\log p(i) = \theta_0 + \sum_{j : S(j) = D, j_D = i_D, D \in \mathcal{D}} \theta_j, \tag{2.6}$$

where θ_0 is the unique number such that $\sum_{i \in I} p(i) = 1$.

Again, to simplify notation, for $i \in I$ and $j \in J$, we write

$$j \lhd i$$

to mean that $S(j)$ is contained in $S(i)$ and that $j_{S(j)} = i_{S(j)}$. Note that we use the symbol $\lhd$ rather than the traditional $\prec$ for partial ordering because $\lhd$ is a partial ordering on J but not on I. We will never use the notation $i \lhd i'$ for i and i' in $I \setminus J$. However $\lhd$ has the property that if $j, j' \in J$ and $i \in I$, then

$$j \lhd j' \quad \text{and} \quad j' \lhd i \Rightarrow j \lhd i. \tag{2.7}$$

Associated to the partial ordering $\lhd$ on J, there are two classical functions on $J \times J$ which will be used in the sequel: the ζ function and the Moebius function μ defined as follows:

$$\zeta(j, j') = 1 \qquad \text{if } j \lhd j' \quad \text{and} \quad 0 \text{ otherwise}; \tag{2.8}$$

$$\mu(j, j') = (-1)^{|S(j')| - |S(j)|} \qquad \text{if } j \lhd j' \quad \text{and} \quad 0 \text{ otherwise}. \tag{2.9}$$

A proof of the fact that (2.9) is indeed the Moebius function of the poset $(J, \lhd)$ is in the proof of Lemma 2.1 of the supplementary file [15]. Using the symbol $\lhd$, representation (2.6) becomes

$$\log p(i) = \theta_0 + \sum_{j \lhd i} \theta_j. \tag{2.10}$$

EXAMPLE 2.1. 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 $a, a, b, b, c, ab, ab, ab, ab, bc, bc$. For $i = 201$ the set of j in J such that $j \lhd i$ is $\{200, 001\}$. For $i = 211$ this set is $\{210, 200, 011, 001, 010\}$ and so on. For these two cells, the unique representation (2.10) for $\log p(i)$ is

$$\log p(201) = \theta_0 + \theta_{200} + \theta_{001},$$
$$\log p(211) = \theta_0 + \theta_{200} + \theta_{010} + \theta_{001} + \theta_{210} + \theta_{011}.$$

We now proceed to give the general matrix form of the loglinear model (2.10). We fix an arbitrary order of the elements of I and of the elements of J. Let $(g_i)_{i \in I}$ and $(e_j)_{j \in J}$ be the canonical basis of $\mathbb{R}^I$ and $\mathbb{R}^J$, respectively, each endowed with their natural Euclidean structure. In our example above, the g_i's are 18-dimensional vectors with components equal to 0 except for the component corresponding to cell $i \in I$ which is 1, while the e_j are 11-dimensional vectors with all components equal to 0 except for that corresponding to the cell $j \in J$. Using the notation

$$\log p = \big(\log p(0), \log p(i), i \in I \setminus \{0\}\big)^t, \qquad \theta = (\theta_j, j \in J)^t$$

and

$$\tilde{\theta} = (\theta_0, \theta_j, j \in J)$$

we have the following.

PROPOSITION 2.1. *The loglinear model defined by the representation* (2.10) *can be written under matrix form as*

$$\log p = X\tilde{\theta}, \tag{2.11}$$

where X is an $(|I|) \times (1 + |J|)$ matrix. Its first column is equal to $\mathbf{1}_I$, the vector with all components equal to 1 in $\mathbb{R}^I$. The other columns are indexed by $j \in J$ and are equal to

$$\sum_{i \in I, j \lhd i} g_i, \qquad j \in J. \tag{2.12}$$

The rows of X are indexed by $i \in I$ and equal to $\tilde{f_i}^t = (1, f_i^t) \in \mathbb{R}^{J+1}$ where

$$f_i = \sum_{j \in J, j \lhd i} e_j \tag{2.13}$$

with $\tilde{f}_0^{\,t} = (1, 0, \ldots, 0)$. *Equivalently* (2.11) *can be written*

$$\left(\log \frac{p(i)}{p(0)}, i \in I \setminus \{0\}\right) = X_{-0}\theta, \tag{2.14}$$

where X_{-0} *is the* $(|I| - 1) \times |J|$ *matrix deduced from* X *by removing its first row and first column. The rows of* X_{-0} *are the* $f_i^t, i \in I$.

The parameter $\theta \in \mathbb{R}^J$ *is uniquely defined by*

$$\theta_j = \sum_{j' \in J; j' \lhd j} (-1)^{|S(j)| - |S(j')|} \log \frac{p(j')}{p(0)}. \tag{2.15}$$

Moreover, the columns of X *form a basis of* $\Omega_{\mathcal{D}}$ *which is therefore of dimension*

$$1 + |J| \qquad \text{with } |J| = \sum_{D \in \mathcal{D}} \prod_{v \in D} (|I_v| - 1). \tag{2.16}$$

Under multinomial sampling, θ_0 *is uniquely defined by* $e^{-\theta_0} = p(0)^{-1} = L(\theta)$, *where*

$$L(\theta) = 1 + \sum_{i \in I \setminus \{0\}} \exp\langle f_i, \theta\rangle = \sum_{i \in I} \exp\langle f_i, \theta\rangle. \tag{2.17}$$

PROOF. The expressions (2.12), (2.13) and (2.14) follow immediately from representation (2.10) and the definitions of $g_i, i \in I$, and $e_j, j \in J$. The $|J| \times |J|$ matrix X_J obtained from X by keeping only the rows and columns indexed by J is representative of the zeta function [see (2.8)] of $\lhd$, the partial order defined above on J. The matrix X_J is therefore invertible, and its columns are independent. The columns of X_{-0} are also therefore independent. The inverse of X_J is given by the Moebius function [see (2.9)] of the partial order on J. So (2.15) follows immediately from the Mobius inversion theorem. Since $\mathbf{1}_I \in \mathbb{R}^{|I|}$ is clearly independent of the other columns of X_{-0} in $\mathbb{R}^{|I|-1}$, the $1 + |J|$ columns of X form a basis of $\Omega_{\mathcal{D}}$. Thus the dimension of $\Omega_{\mathcal{D}}$ is given by (2.16). This dimension is also given in [4] and [13]. To prove (2.17), we need only observe that $\frac{p(i)}{p(0)} = e^{\langle\theta, f_i\rangle}, i \in I \setminus \{0\}$, and $p(0) = 1 - \sum_{i \in I \setminus \{0\}} p(i)$ and solve for $p(0)$. □

EXAMPLE 2.2. For the model defined by $V = \{a, b, c\}$, $\mathcal{D} = \{a, b, c, ab, bc\}$ and $I_a = \{0, 1\} = I_b = I_c$, we have $I = (000, 100, 010, 110, 001, 101, 011, 111)$ and $J = \{(100), (010), (001), (110), (011)\}$. Then

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \begin{array}{l} f_{000} = (0, 0, 0, 0, 0)^t, \\ f_{100} = (1, 0, 0, 0, 0)^t, \\ f_{010} = (0, 1, 0, 0, 0)^t, \\ f_{110} = (1, 1, 0, 1, 0)^t, \\ f_{001} = (0, 0, 1, 0, 0)^t, \\ f_{101} = (1, 0, 1, 0, 0)^t, \\ f_{011} = (0, 1, 1, 0, 1)^t, \\ f_{111} = (1, 1, 1, 1, 1)^t. \end{array}$$

We also have

$$X_J = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad X_J^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \end{pmatrix}.$$

As mentioned above, our presentation (2.10) of the hierarchical loglinear model defined by $\Omega_{\mathcal{D}}$ relies on the characterization of this model in Proposition B.4 and formula (B.11) of [14]; see also [4]. We offer a different proof of this characterization in Section 2.1 of the supplementary file [15].

2.2. *The multinomial distribution as a natural exponential family.* We consider a contingency table with cells $i = (i_v, v \in V) \in I$ and cell counts $n = (n(i), i \in I)$ with $\sum_{i \in I} n(i) = N$ obtained from N i.i.d. observations of a multivariate Bernoulli variable with parameter $(p(i), i \in I)$, that is, with distribution $\sum_{i \in \mathcal{I}} p(i)\delta_{g_i}$. For $E \subset V$ we write $n_E(i_E) = \sum_{i' \in I; i_E = i'_E} n(i')$ for the E-marginal count. For the particular case $E = S(j)$, $j \in J$, we write

$$t(j) = n_E(j_E). \tag{2.18}$$

Then, using (2.11), we have

$$\sum_{i \in I} n(i) \log p(i) = \langle \log p, n \rangle_{\mathbb{R}^I} = \langle X\tilde{\theta}, n \rangle = \langle \tilde{\theta}, X^t n \rangle, \tag{2.19}$$

where, from (2.12), $X^t n = (\sum_{i \in I} n(i), \sum_{j \lhd i} n(i), j \in J) = (N, t(j), j \in J)$. We therefore have

$$\sum_{i \in I} n(i) \log p(i) = N\theta_0 + \sum_{j \in J} t(j)\theta_j \tag{2.20}$$

and, using (2.17)

$$\begin{aligned} \prod_{i \in I} p(i)^{n(i)} &= \exp\Big(\sum_{j \in J} t(j)\theta_j - N \log\Big(\sum_{i \in I} \exp\langle f_i, \theta \rangle\Big)\Big) \\ &= \frac{\exp \sum_{j \in J} t(j)\theta_j}{L(\theta)^N}. \end{aligned} \tag{2.21}$$

EXAMPLE 2.3. For Example 2.2, the vector $t_J = (t(j), j \in J)$ of sufficient statistics is

$$\begin{aligned} &(t(100), t(010), t(001), t(110), t(011)) \\ &\quad = (n_a(1), n_b(1), n_c(1), n_{ab}(1, 1), n_{bc}(1, 1)). \end{aligned}$$

The multinomial distribution for the model generated by $\mathcal{D}$ is therefore a natural exponential family on $\mathbb{R}^J$ characterized by the set J defined in (2.4). The family is generated by a discrete measure on $\mathbb{R}^J$ whose Laplace transform is $L(\theta)^N = (\sum_{i\in I} e^{\langle\theta,f_i\rangle})^N$. Clearly L is the Laplace transform of the counting measure

$$\mu = \sum_{i\in I} \delta_{f_i} \tag{2.22}$$

on the set of vectors $(f_i)_{i\in I}$. This exponential family is concentrated on a bounded set of $\mathbb{R}^J$, and therefore the set of parameters θ for which L is finite is the whole space $\mathbb{R}^J$. Hence the family is regular in the sense of Barndorff-Nielsen [2] and Diaconis and Ylvisaker [7]. Since f_0 is the zero vector in $\mathbb{R}^J$ and $(N, t(j), j \in J) = X^t n = \sum_{i\in I} n(i)\tilde{f_i}$, from (2.19), (2.20) and (2.21) it is clear that the vector of sufficient statistics

$$\frac{t_J}{N} = \left(\frac{t(j)}{N}, j \in J\right)^t = \sum_{i\in I\setminus\{0\}} \frac{n(i)}{N} f_i = \sum_{i\in I} \frac{n(i)}{N} f_i \tag{2.23}$$

belongs to the convex hull of $(f_i)_{i\in I}$. Let $C \subset \mathbb{R}^J$ be the interior of this convex hull. In Proposition 2.2 below we state that the $(f_i)'s$ are the extreme points of its closure $\overline{C}$. The proof is given in the supplementary file [15].

PROPOSITION 2.2. *The extreme points of the convex hull of the support of the measure μ as defined in* (2.22) *are the $f_i, i \in I$, as defined in* (2.13).

2.3. *The DY conjugate prior for the loglinear parameters*. From the form (2.21) of the multinomial distribution and Theorem 1 in [7], the DY conjugate prior distribution for θ has density with respect to the Lebesgue measure equal to

$$\pi(\theta|m_J, \alpha, J) = \frac{1}{I_J(m_J, \alpha)} \times \frac{e^{\alpha\langle\theta,m_J\rangle}}{L(\theta)^\alpha},$$

where $I_J(m, \alpha)$ is the normalizing constant. It is proper if and only if the hyper-parameter (α, m_J) is such $\alpha > 0$ and $m_J \in C$. The posterior probability of θ given the data $n = (n(i))_{i\in I}$ and t_J as defined in (2.23) is

$$\pi\left(\theta\middle|\frac{\alpha m_J + t_J}{\alpha + N}, \alpha + N, J\right).$$

In classical Bayesian model selection, the most probable models are selected by means of Bayes factors. More precisely, models are compared two by two by means of the Bayes factor $B_{1,2}$ between model J_1 and model J_2. If the prior on the set of all hierarchical models is uniform, we have

$$B_{1,2} = \frac{I_2(m_2, \alpha)}{I_1(m_1, \alpha)} \times \frac{I_1((\alpha m_1 + t_1)/(\alpha + N), \alpha + N)}{I_2((\alpha m_2 + t_2)/(\alpha + N), \alpha + N)}, \tag{2.24}$$

where, for the sake of simplicity, m, t, I are indexed by $k = 1, 2$ rather than by J_1, J_2 and where m_1 and m_2 have been chosen in C_1 and C_2, respectively. The aim of the present paper is to find the limit of $B_{1,2}$ when $\alpha \to 0$. If we assume that $n(i) > 0$ for all $i \in I$, then t_k/N is in the interior of C_k, and under these circumstances the second factor in the right-hand side of (2.24) has the finite limit $I_1(\frac{t_1}{N}, N)/I_2(\frac{t_2}{N}, N)$. For the first factor in (2.24), we will show that $I(m, \alpha) \sim_{\alpha\to 0} \mathbb{J}_C(m)\alpha^{-|J|}$ where $\mathbb{J}_C(m)$ will be defined in the next section. Thus when $\alpha \to 0$ the Bayes factor is equivalent to

$$\alpha^{|J_1|-|J_2|}\frac{\mathbb{J}_{C_2}(m_2)}{\mathbb{J}_{C_1}(m_1)} \times \frac{I_1(t_1/N, N)}{I_2(t_2/N, N)}.$$

If we do not assume that $n(i) > 0$ for all $i \in I$, then t_k/N might be on the boundary of C_k for at least one $k = 1, 2$ and we will have to further study the behavior of $I(m, \alpha)$ and $\mathbb{J}_C(m)$. This is done in the following section.

3. The limiting behavior of the prior normalizing constant. We give three fundamental theoretical results in this section. We assume that m is in the interior of C, the convex hull of the measure μ as defined in (2.22). Theorem 3.1 gives the general form of $\mathbb{J}_C(m)$ in terms of the affine forms defining the facets of C. Theorem 3.2 gives the limit of $I(m, \alpha)$ when $\alpha \to 0$, and Theorem 3.3 describes the behavior of $\mathbb{J}_C(\lambda m + (1-\lambda)y)$ when $y \in \overline{C} \setminus C$ and $\lambda \to 0$.

3.1. *The characteristic function of a convex set.* Given a finite-dimensional real linear space E, let E^* be its dual, that is, the space of all linear forms θ on E. We write $\langle\theta, x\rangle$ instead of $\theta(x)$ when $(\theta, x) \in E^* \times E$. We fix a Lebesgue measure $d\theta$ on E^* and a Lebesgue measure dx on E which must be compatible (this means that if e is a basis of E, and e^* is the corresponding dual basis of E^*, the product of the respective volumes of the two cubes built on e and e^* must be one). Needless to say when $E = \mathbb{R}^n$, then $E^* = E$, $\langle\cdot,\cdot\rangle$ is the usual inner product, and the Lebesgue measure is the usual one. It will, however, be important in the sequel to distinguish between E and E^*, and we therefore keep this notation.

If $C \subset E$ is an open nonempty convex set not containing an (affine) line, its polar set is

$$C^o = \{\theta \in E^*; \langle\theta, x\rangle \le 1\ \forall x \in C\},$$

its support function $h_C : E^* \to (-\infty, \infty]$ is

$$h_C(\theta) = \sup\{\langle\theta, x\rangle; x \in C\}$$

and its characteristic function is the function $m \mapsto \mathbb{J}_C(m)$ defined on C by

$$\mathbb{J}_C(m) = \int_{E^*} e^{\langle\theta,m\rangle - h_C(\theta)}\, d\theta. \tag{3.1}$$

We note that if C contained a line, we would have $h_C(\theta) = \infty$ almost everywhere and $\mathbb{J}_C \equiv 0$. Faraut and Koranyi ([10], page 10) define $\mathbb{J}_C$ when C is an open convex salient cone. In that case, the polar set of C is the convex cone

$$C^o = \{\theta \in E^*; \langle \theta, x\rangle \le 0\ \forall x \in C\}, \tag{3.2}$$

and $h_C(\theta) = 0$ if $\theta \in C^o$ and $h_C(\theta) = \infty$ if $\theta \notin C^o$. When C is a bounded set, $h_C(\theta)$ is finite for all $\theta \in E^*$. We also have the following important property of $\mathbb{J}_C(\cdot)$. Its proof can be found in the supplementary file [15].

LEMMA 3.1. *Let C be an open convex set not containing a line, and let $m \in C$. Then $\mathbb{J}_C(m)$ is finite.*

One can prove that $\mathbb{J}_C(m) = \infty$ if $m \notin C$. Another property of $\mathbb{J}_C(m)$ is that when C is an open convex set of $\mathbb{R}^n$ not containing a line, the following formulas hold:

$$\mathbb{J}_C(m) = n!\,\mathrm{Vol}(C - m)^o = n! \int_{C^o} \frac{d\theta}{(1 - \langle \theta, m\rangle)^{n+1}}. \tag{3.3}$$

For the first equality in (3.3), see [3], page 207, and [1], page 243. For the second one, make the change of variable $\theta = \theta'/(1 + \langle \theta', m\rangle)$ in the integral $\int_{(C-m)^o} d\theta'$. Computing $\mathbb{J}_C(m)$ when C is associated to an arbitrary hierarchical model is usually difficult except, as we shall see in Section 5.2, when the model is a graphical decomposable model. Consider, however, the following simple example:

EXAMPLE 3.1. Let $C = (0, 1) \subseteq \mathbb{R}$. In this case, $h_C(\theta) = \max(0, \theta)$, and for $0 < m < 1$, we have

$$\mathbb{J}_C(m) = \int_{-\infty}^{0} e^{\theta m}\, d\theta + \int_{0}^{\infty} e^{\theta m - \theta}\, d\theta = \frac{1}{m} + \frac{1}{1-m} = \frac{1}{m(1-m)}. \tag{3.4}$$

Two more examples of $\mathbb{J}_C(m)$ will be given after Theorem 3.2 below. We now give a theorem that states that $\mathbb{J}_C(m)$ is the ratio of polynomials where the denominator is equal to the product of the affine forms defining the facets of $\overline{C}$. This will be used in Section 5 to identify the facets of $\overline{C}$ for decomposable graphical models. We first need the following lemma which computes the characteristic function of a simplicial cone.

LEMMA 3.2. *Let $(x_1, \ldots, x_n)$ be a basis of E, and let $(\xi_1, \ldots, \xi_n)$ be its dual basis in E^* (i.e., $\langle \xi_j, x_i\rangle = \delta_i^j$). Consider the simplicial cone A of E^* defined by*

$$\begin{aligned} A &= \{\theta = \theta_1\xi_1 + \cdots + \theta_n\xi_n; \theta_1 > 0, \ldots, \theta_n > 0\} \\ &= \{\theta \in E^*; \langle \theta, x_1\rangle > 0, \ldots, \langle \theta, x_n\rangle > 0\}, \end{aligned}$$

and denote by $\mathrm{Vol}(\xi_1, \dots, \xi_n)$ *the volume of the parallelotope*

$$\{\theta = \theta_1\xi_1 + \cdots + \theta_n\xi_n; 0 \le \theta_1 \le 1, \dots, 0 \le \theta_n \le 1\}.$$

Then for all x *in* $-A^o \subset E$, *that is*, *the opposite of the dual cone of* A, *we have*

$$\int_A e^{-\langle\theta,x\rangle}\, d\theta = \frac{\mathrm{Vol}(\xi_1, \dots, \xi_n)}{\langle\xi_1, x\rangle \cdots \langle\xi_n, x\rangle}.$$

This lemma is elementary and is obtained by writing θ in the ξ basis and by making the change of variable from the coordinates of θ in the canonical basis of $\mathbb{R}^n$ to the coordinates in the ξ basis.

Recall that a *facet* of a polytope $\overline{C} \subset \mathbb{R}^n$ with a nonempty interior is a face of dimension $n-1$. More specifically a facet is the intersection of $\overline{C}$ with a supporting hyperplane of $\overline{C}$ which contains n affinely independent points of $\overline{C}$.

THEOREM 3.1. *Let* $C \subset E$ *be the nonempty interior of a bounded polytope* $\overline{C}$. *Let* $m \in C$. *Then we have*

$$\mathbb{J}_C(m) = \frac{N(m)}{D(m)},$$

where $D(m) = \prod_{k=1}^K g_k(m)$ *is the product of affine forms* $g_k(m)$ *in* m *such that* $g_k(m) = 0, k = 1, \dots, K$, *define the facets of* $\overline{C}$ *and where* $N(m)$ *is a polynomial of degree* $< K$.

The proof is in the supplementary file [15]. The idea of the proof is to partition the integrating space E^* into the cones $A(f)$ dual to the supporting cones to C at f for $f \in \{f_i, i \in I\}$. Each $A(f)$ is in turn split into a sum of simplicial cones, and Lemma 3.2 is then used to compute these integrals.

3.2. *The behavior of* $I(m, \alpha)$ *as* $\alpha \to 0$. We have the following theorem.

THEOREM 3.2. *Let* μ *be a positive measure on the* n-*dimensional linear space* E *with closed convex support bounded and with nonempty interior* C. *Denote by* $L(\theta) = \int_E e^{\langle\theta,x\rangle}\mu(dx)$ *its Laplace transform. For* $m \in C$ *and for* $\alpha > 0$ *consider the Diaconis–Ylvisaker integral*,

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

Then

$$\lim_{\alpha\to 0} \alpha^n I(m, \alpha) = \mathbb{J}_C(m). \tag{3.5}$$

Let us note immediately that a remarkable feature of this result is that the limit $\mathbb{J}_C(m)$ of $\alpha^n I(m,\alpha)$ depends on μ only through its convex support. For instance, if $E=\mathbb{R}$, the uniform measure on $(0,1)$ and the sum $\mu=\delta_0+\delta_1$ of two Dirac measures share the same $C=(0,1)$ and the same $\mathbb{J}_C(m)=(m(1-m))^{-1}$. We now need the following lemma.

LEMMA 3.3. *Let μ be a bounded measure on some measurable space Ω and let f be a positive, bounded and measurable function on Ω. Then we have*:

(1) $\|f\|_p \to_{p\to\infty} \|f\|_\infty$;
(2) *The function $p\mapsto \|f\|_p$ is either decreasing on $(0,\infty)$ or there exists $p_0\geq 0$ such that it is decreasing on $(0,p_0]$ and increasing on $[p_0,+\infty)$.*

The proof of this lemma is simple and can be found in the supplementary file [15].

PROOF OF THEOREM 3.2. In the integral $\alpha^n I(m,\alpha)$ we make the change of variable $y=\alpha\theta$, and we obtain

$$\alpha^n I(m,\alpha)=\int_{E^*}\frac{e^{\langle y,m\rangle}}{L(y/\alpha)^\alpha}\,dy.$$

We now apply the last lemma to $\Omega=\overline{C}$, to the bounded measure μ, to the function $f(x)=e^{\langle y,x\rangle}$ for some fixed $y\in E^*$ and to $p=1/\alpha$. Denote by S the support of μ. One easily sees that the support function of C satisfies

$$h_C(\theta)=\sup\{\langle\theta,x\rangle; x\in C\}=\max\{\langle\theta,x\rangle; x\in S\}$$

since C is the interior of the convex hull of S. As a consequence the essential sup of f is $e^{h_C(y)}$ and we get $\lim_{\alpha\to 0}L(y/\alpha)^\alpha=e^{h_C(y)}$. Furthermore, by Lemma 3.3, the function $p\mapsto\|f\|_p$ is monotonic for p big enough. If $p\mapsto\|f\|_p$ is increasing, $\frac{1}{\|f\|_p}$ is decreasing, and then by the monotone convergence theorem,

$$\lim_{\alpha\to 0}\int_{E^*}\frac{e^{\langle y,m\rangle}}{L(y/\alpha)^\alpha}\,dy=\int_{E^*}\frac{e^{\langle y,m\rangle}}{\lim_{\alpha\to 0}L(y/\alpha)^\alpha}\,dy=\int_{E^*}e^{\langle y,m\rangle-h_C(y)}\,dy=\mathbb{J}_C(m).$$

If $p\mapsto\|f\|_p$ is decreasing, $p\mapsto 1/\|f\|_p$ is increasing. In order to show that we can invert the order of limit and integration and apply the monotone convergence theorem as we did in the previous case, we need to insure that $\int_{E^*}e^{\langle y,m\rangle-h_C(y)}\,dy$ is finite: Lemma 3.1 shows that it is true. □

We now give two more examples of functions $\mathbb{J}_C(m)$ which we compute using Theorem 3.2.

EXAMPLE 3.2. Let $e_0 = 0$ and $(e_1, \dots, e_n)$ be the canonical basis of $\mathbb{R}^n$. Let C be the interior of the simplex generated by $e_0, \dots, e_n$. Then C is the set of $m \in \mathbb{R}^n$ such that $m = \sum_{j=0}^n \lambda_j e_j$ for some unique positive $\lambda_0, \dots, \lambda_n$ satisfying $\lambda_1 + \cdots + \lambda_n < 1$. In this case

$$J_C(m) = \frac{1}{m_1 m_2 \cdots m_n (1 - m_1 - \cdots - m_n)}.$$

This result can be obtained by computing $I(m, \alpha)$ for $\mu = \delta_{e_0} + \sum_{i=1}^n \delta_{e_i}$. Using elementary methods of integration, we find that

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

where $m_0 = 1 - \sum_{i=1}^n m_i$. Using $z\Gamma(z) = \Gamma(1 + z) \to_{z\to 0} = 1$ we immediately obtain that

$$\mathbb{J}_C(m) = \lim_{\alpha\to 0} \alpha^n I(m, \alpha) = \frac{1}{\prod_{i=0}^n m_i}.$$

EXAMPLE 3.3. Consider the graphical model with decomposable graph $\overset{a}{\bullet} - \overset{b}{\bullet} - \overset{c}{\bullet}$. For simplicity, we will assume that the variables a, b, c are binary so that $m = (m_j, j \in J)$ can be written $m = (m_D, D \in \mathcal{D})$ where $\mathcal{D} = \{a, b, c, ab, bc\}$. We shall generalize this example in Section 5. From formula (4.8) in [16], we know that

$$\begin{aligned} I(m, \alpha) = {} & \Gamma\bigl(\alpha(1 - m_a - m_b + m_{ab})\bigr)\Gamma\bigl(\alpha(m_a - m_{ab})\bigr) \\ & \times \Gamma\bigl(\alpha(m_b - m_{ab})\bigr)\Gamma(\alpha(m_{ab}))\Gamma\bigl(\alpha(1 - m_b - m_c + m_{bc})\bigr) \\ & \times \Gamma\bigl(\alpha(m_b - m_{bc})\bigr)\Gamma\bigl(\alpha(m_c - m_{bc})\bigr)\Gamma(\alpha(m_{bc})) \\ & \times \frac{1}{\Gamma(\alpha m_b)\Gamma(\alpha(1 - m_b))}, \end{aligned}$$

and therefore using $z\Gamma(z) = \Gamma(1 + z) \to_{z\to 0} 1$ again we obtain that

$$\begin{aligned} \lim_{\alpha\to 0} \alpha^5 I(m, \alpha) &= \mathbb{J}_C(m) \\ &= \frac{m_b(1 - m_b)}{m_{ab} m_{bc}} \\ &\quad \times \frac{1}{(1 - m_a - m_b + m_{ab})(m_a - m_{ab})(m_b - m_{ab})} \\ &\quad \times \frac{1}{(1 - m_b - m_c + m_{bc})(m_b - m_{bc})(m_c - m_{bc})}. \end{aligned}$$

3.3. *The behavior of* $\mathbb{J}_C(\lambda m + (1-\lambda)y)$ *when* $y \in \overline{C} \setminus C$ *and* $\lambda \to 0$. In practice, the choice of the hyperparameters m and α is ours, and for a given model J, it is traditional to take $m = (m_j, j \in J)$ to be the vector of J-marginal counts in a fictive contingency table with cell counts all equal and equal to $\frac{1}{|I|}$. In any case, as long as all fictive cell counts are positive, m belongs to the open set C and the behavior of $I(m, \alpha)$ is given by Theorem 3.2. When studying the Bayes factor, we will have to consider the case where the data belongs to the boundary $\overline{C} \setminus C = \partial C$ of C, that is to a face of $\overline{C}$. To do so, we will need to describe the behavior of $\mathbb{J}_C(z)$ as z approaches the boundary of C along a straight line. This is done in the following theorem.

THEOREM 3.3. *Let* $C \subset E$ *be an open polytope with* $\dim E = n$. *Let* $y \in \partial C$, *let* F *be the face of* $\overline{C}$ *containing* y *in its relative interior and let* k *be the dimension of* F. *Then when* $\lambda \to 0$,

$$\lim_{\lambda \to 0} \lambda^{n-k} \mathbb{J}_C\bigl(\lambda m + (1-\lambda)y\bigr) = D,$$

where D *is a positive constant.*

The proof is in the Appendix.

4. The limiting behavior of the Bayes factor.

Let us recall that, under the uniform distribution on the class of hierarchical models, the Bayes factor between two models J_1 and J_2 is equal to

$$B_{1,2} = \frac{I_1((\alpha m_1 + t_1)/(\alpha + N), \alpha + N) I_2(m_2, \alpha)}{I_2((\alpha m_2 + t_2)/(\alpha + N), \alpha + N) I_1(m_1, \alpha)},$$

where $t_i = t_{J_i} = (t(j), j \in J_i), i = 1, 2$. The central result of this section is Corollary 4.2 which gives the behavior of $B_{1,2}$ depending on where the data $\frac{t_i}{N}$ sits on $\overline{C}_i, i = 1, 2$. This result covers all possible cases. The first possible case is that both $\frac{t_i}{N}$ are in C_i. In that case, each data point is on the face of C_i of dimension $k_i = |J_i|$. In the second case, we have $\frac{t_1}{N}$ in C_1, that is, on the face of dimension $k_1 = |J_1|$, and $\frac{t_2}{N}$ in $\overline{C}_2 \setminus C_2$ on a face of dimension $k_2 < |J_2|$. Similarly if we have $\frac{t_1}{N} \in \overline{C}_1 \setminus C_1$ and $\frac{t_2}{N} \in C_2$. In the third case, we have both $\frac{t_i}{N} \in \overline{C}_i \setminus C_i$ on faces of dimension $k_i < |J_i|$, respectively. For the first case, as we already know, we need only look at the behavior of $I(m, \alpha)$ when $\alpha \to 0$ and the answer is given by Theorem 3.2. We consider this case in Section 4.1. For the second and third cases, we need to look at $I(m, \alpha)$ and also at $I(\frac{\alpha m_i + t_i}{\alpha + N}, \alpha + N)$ when $\alpha \to 0$. This is done in Theorem 4.1.

4.1. *The case where the data is in the interior C of $\overline{C}$.* The data is given in the form of a contingency table with cell counts $n = (n(i), i \in I)$. We consider now the case where the data, which appears under the form t_i in models J_i, belongs to C_i, $i = 1, 2$, so that $I_i(\frac{t_i}{N}, N)$, $i = 1, 2$, are finite and positive. In this case, as $\alpha \to 0$, from Theorem 3.2, we know that

$$B_{1,2} \sim \alpha^{|J_1|-|J_2|} \frac{I_1(t_1/N, N)\mathbb{J}_{C_2}(m_2)}{I_2(t_2/N, N)\mathbb{J}_{C_1}(m_1)}, \tag{4.1}$$

where we recall that $|J_i| = \dim C_i$. Since the numbers $\mathbb{J}_{C_i}(m_i)$, $i = 1, 2$, are finite and positive, we have the following corollary of Theorem 3.2.

COROLLARY 4.1. *When the data belong to the open polytope C_i, $i = 1, 2$, the Bayes factor $B_{1,2}$ is such that, when $\alpha \to 0$,*

$$B_{1,2} \sim \alpha^{|J_1|-|J_2|}.$$

This implies in particular that, when the data is in both C_i, $i = 1, 2$, the Bayes factor always favors the sparser model.

The proof follows immediately from (4.1). Moreover, when $\alpha \to 0$ and $|J_2| < |J_1|$, $B_{1,2}$ tends to 0. This result has been well known, at least numerically, for the class of decomposable models, and in that case, it can be proved by expressing the Bayes factor as in (4.8) of [16] and using the fact that $\Gamma(\alpha) \sim \alpha^{-1}$ as $\alpha \to 0$; see Example 3 of Section 3 and Section 5.2. It has also been observed to hold numerically, most of the time, for hierarchical models. Computations illustrating the fact that the Bayes factor tends to favor the sparser models in the class of all hierarchical models can be found in [16], page 3456. We have just shown that it always holds when the data is in C_1 and in C_2. We will see in the next subsection that things are more delicate when the data belongs to the boundary of at least one of $\overline{C}_1$ or $\overline{C}_2$.

4.2. *The case where the data belongs to a face of $\overline{C}_i$, $i = 1, 2$.* When $\alpha \to 0$, $\frac{\alpha m_i + t_i}{\alpha + N}$ converges to the boundary point $\frac{t_i}{N}$ of C_i along the segment

$$s(\alpha) = \frac{\alpha m_i + t_i}{\alpha + N} = \frac{\alpha}{\alpha + N} m_i + \left(1 - \frac{\alpha}{\alpha + N}\right)\frac{t_i}{N}. \tag{4.2}$$

We need to study the limiting behavior of $B_{1,2}$ when $\alpha \to 0$. To do so, we will use Theorem 3.3 to obtain the following result.

THEOREM 4.1. *Suppose that $\frac{t}{N} \in \overline{C} \setminus C$ belongs to the relative interior of a face F of dimension k. Then*

$$\lim_{\alpha \to 0} \alpha^{(|J|-k)} I\left(\frac{\alpha m + t}{\alpha + N}, \alpha + N\right) \tag{4.3}$$

exists and is positive.

The proof of Theorem 4.1 is given in the supplementary file [15]. From Theorems 3.2 and 4.1, we immediately derive the following which is the object of this paper.

COROLLARY 4.2. *Consider two hierarchical models $J_i, i = 1, 2$, of dimension $|J_i|$. Assume that the data $\frac{t_i}{N}$ belongs to the relative interior of a face F_i of C_i of dimension k_i. Then the asymptotic behavior of the Bayes factor $B_{1,2}$ when $\alpha \to 0$ is given by*

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

where D is a finite positive constant. The Bayes factor favors the model which contains the data in the relative interior of the face of C_i of smallest dimension.

The proof is immediate. According to Theorems 3.2 and 4.1, we have

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

REMARK 4.1. We note that, if $\frac{t_i}{N} \in C_i, i = 1, 2$, since C_i is the face of $\overline{C}_i$ of dimension J_i, then $k_i = |J_i|$ and Corollary 4.2 yields Corollary 4.1. For the same reason, Corollary 4.2 also deals with the cases where $\frac{t_i}{N} \in C_i$ for only one of $i = 1$ or $i = 2$.

4.3. *The results of Steck and Jaakola* [20] *as a particular case.* In [20] Steck and Jaakola study the behavior of the Bayes factor for two Bayesian network models differing by one edge only, when $\alpha \to 0$. They show it is equivalent to the problem of comparing two Bayesian network models with three variables indexed by $\{a, b, c\}$. The first model has directed edges (b, a), (b, c) and (a, c). The second model has directed edges (b, a) and (b, c). These two Bayesian network models are Markov equivalent to the two hierarchical (in fact graphical) models J_1 and J_2 with, respectively, generating sets $\mathcal{D}_1 = \{abc\}$ and $\mathcal{D}_2 = \{ab, bc\}$. Moreover on these two models, the prior in [20] is equivalent to ours. We must then be able to compare their result given in Proposition 1 of [20] and our result given in Corollary 4.2. To give their results Steck and Jaakola [20] introduce the quantity

$$d_{\mathrm{EDF}} = \sum_{i \in \mathcal{I}} \delta(n(i)) - \sum_{i_{ab} \in \mathcal{I}_{ab}} \delta(n(i_{ab})) - \sum_{i_{bc} \in \mathcal{I}_{bc}} \delta(n_{i_{bc}}) + \sum_{i_b \in \mathcal{I}_b} \delta(n(i_b)), \tag{4.4}$$

where $\delta(\cdot)$ is an indicator function which is such that $\delta(x) = 0$ if $x = 0$ and $\delta(x) = 1$ otherwise. They state that the Bayes factor $B_{1,2}$ behaves as follows:

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

This result coincides with our Corollaries 4.1 and 4.2 for three variable models. In fact, we are going to show the following.

PROPOSITION 4.1. *Consider the two decomposable graphical models on three variables, J_1 and J_2, as defined above. If the data belongs to faces of dimension k_1 and k_2 of, respectively, C_1 and C_2, then we have*

$$d_{\mathrm{EDF}} = k_1 - k_2.$$

PROOF. The Bayes factor is equal to

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

where the form of the normalizing constants $I(m, \alpha)$ for decomposable models is well known; see, for example, equation (4.8) of [16]. When $\alpha \to 0$, from Theorem 3.2, we know that $I(m_2, \alpha)/I(m_1, \alpha) \sim \alpha^{|J_1|-|J_2|}$.

Expressed in terms of cell counts for the full table, for the b-, ab- and bc-marginal tables, we have

$$\begin{aligned}
&\frac{I((\alpha m_1 + t_1)/(\alpha + N), \alpha + N)}{I((\alpha m_2 + t_2)/(\alpha + N), \alpha + N)} \\
&\quad = \frac{\prod_{i \in \mathcal{I}} \Gamma(\alpha m(i) + n(i)) \prod_{i_b \in \mathcal{I}_b} \Gamma(\alpha m(i_b) + n(i_b))}{\prod_{i_{ab} \in \mathcal{I}_{ab}} \Gamma(\alpha m(i_{ab}) + n(i_{ab})) \prod_{i_{bc} \in \mathcal{I}_{bc}} \Gamma(\alpha m(i_{bc}) + n(i_{bc}))}.
\end{aligned}$$

If for some $D = \varnothing, ab, bc, b$, the marginal cell count $n(i_D)$ is different from 0, when $\alpha \to 0$, $\Gamma(\alpha m(i_D) + n(i_D)) \to \Gamma(n(i_D))$ which is finite. If $n(i_D) = 0$, then $\Gamma(\alpha m(i_D) + n(i_D)) \sim \frac{1}{\alpha m(i_D)}$. It follows that, when $\alpha \to 0$, $B_{1,2} \sim \alpha^q$ where

$$\begin{aligned}
q = {} & \Bigl[|J_1| - \sum_{i \in \mathcal{I}} (1 - \delta(n(i)))\Bigr] \\
& - \Bigl[|J_2| - \sum_{i \in \mathcal{I}_{ab}} \bigl(1 - \delta(n(i_{ab}))\bigr) \\
& - \sum_{i_{bc} \in \mathcal{I}_{bc}} \bigl(1 - \delta(n(i_{bc}))\bigr) + \sum_{i_b \in \mathcal{I}_b} \bigl(1 - \delta(n(i_b))\bigr)\Bigr].
\end{aligned}$$

Let $C_i, i = 1, 2$, be the interior of the convex hull corresponding to model J_i. Consider model J_1 first. It is immediate to see that, following the notation of (5.2) and (5.3) in Section 5 below,

$$\begin{aligned}
n(000) &= g_{0,C_1}, \\
n(i) &= g_{i,C_1}, \qquad i \in \mathcal{I},
\end{aligned}$$

and according to Theorem 5.1, $n(000) = 0$ and $n(i) = 0$ are the equations of the facets of the polytope C_1. Therefore the dimension of the space minus the number

AOS imspdf v.2012/03/16 Prn:2012/03/29; 11:11 F:aos974.tex; (Laima) p. 19

of distinct facets the data belongs to, is equal to the dimension of the face of $\overline{C}_1$ containing the data, that is,

$$|J_1| - \sum_{i\in\mathcal{I}}\bigl(1-\delta(n(i))\bigr) = \sum_{i\in\mathcal{I}}\delta(n(i)) = k_1. \tag{4.5}$$

Similarly, for model J_2, according to Theorem 5.1, the equations of the facets of $\overline{C}_2$ are given by

$$n(i_{ab})=0, \qquad i_{ab}\in\mathcal{I}_{ab}, \quad \text{and} \quad n(i_{bc})=0, \qquad i_{bc}\in\mathcal{I}_{bc}.$$

The facets containing the data are therefore those defined by $n(i_{ab})=0$ or $n(i_{bc})=0$. This does not mean, however, that

$$|J_2| - \biggl(1-\sum_{i_{ab}\in\mathcal{I}_{ab}}\delta(n(i_{ab}))\biggr) - \sum_{i_{bc}\in\mathcal{I}_{bc}}\bigl(1-\delta(n(i_{bc}))\bigr)$$

represents the dimension of the face containing the data. Indeed, if for some $i_b^0\in\mathcal{I}_b$, we have $n(i_b^0)=0$, this means that $n(i_{ab})=0$; also whenever $i_b=i_b^0$ and also $n(i_{bc})=0$ whenever $i_b=i_b^0$. Then, clearly, one of the equations $n(i_{ab})=0$ or $n(i_{bc})=0$ is redundant, and we subtract $1-\delta(n(i_b^0))$ for the count of facets defining the position of the data. It is clear then that

$$|J_2| - \sum_{i\in\mathcal{I}_{ab}}\bigl(1-\delta(n(i_{ab}))\bigr) - \sum_{i_{bc}\in\mathcal{I}_{bc}}\bigl(1-\delta(n(i_{bc}))\bigr) + \sum_{i_b\in\mathcal{I}_b}\bigl(1-\delta(n(i_b))\bigr) = k_2,$$

which, together with (4.5) proves the proposition. $\square$

In fact Proposition 4.1 can be extended to the following general result. Let $\mathcal{C}_i$ and $\mathcal{S}_i$ be the set of cliques and separators of the decomposable model J_i, $i=1,2$. We define the effective degree of freedom to be the following sum d_{EDF}:

$$\begin{aligned} d_{\mathrm{EDF}} = {}& \sum_{C\in\mathcal{C}_1}\sum_{i_C\in\mathcal{I}_C}\delta(n(i_C)) - \sum_{S\in\mathcal{S}_1}\sum_{i_S\in\mathcal{I}_S}\delta(n(i_S)) \\ & - \biggl(\sum_{C\in\mathcal{C}_2}\sum_{i_C\in\mathcal{I}_C}\delta(n(i_C)) - \sum_{S\in\mathcal{S}_2}\sum_{i_S\in\mathcal{I}_S}\delta(n(i_S))\biggr). \end{aligned}$$

PROPOSITION 4.2. *Consider two arbitrary decomposable graphical models J_1 and J_2 such that the data belongs to faces of dimension k_1 and k_2 of C_1 and C_2, respectively. Then, the following relation holds*:

$$d_{\mathrm{EDF}} = k_1 - k_2.$$

The proof of this proposition follows parallel lines to the proof given above. We therefore have a quick and easy way to know the behavior of the Bayes factor between two decomposable models.

5. Facets of $\overline{C}$ for some hierarchical models. From our main result, Corollary 4.2, we see that the behavior of the Bayes factor between two models J_1 and J_2, as $\alpha \to 0$, is determined not only by the specification of the two models but by the position of the data with respect to the support C_i of the multinomial distribution of the model J_i, $i = 1, 2$, respectively. If the data belongs to a face F_i of dimension $k_i < |J_i|$ of C_i, Corollary 4.2 tells us that we ought to consider not the model J_i but the reduced model with support F_i and with dimension k_i. In order to use the Bayes factor correctly for model selection for α small, it is therefore crucial to know which face of C_i the data belongs to. Faces are the intersection of a certain number of facets. So, we must be able to identify the facets of C. This is generally not an easy task.

Facets of the polytope $\overline{C}$ have been much studied by geometers, and in Section 5.3, we will recall some known results on these facets when the model is binary and governed by a cycle of order $n \geq 3$. But before doing so, we give two new results on facets of polytopes associated to our models. In Theorem 5.1, we identify a category of facets which is common to all discrete hierarchical models. In Corollary 5.1, we show that for decomposable graphical models, the only facets of C are given by the category of facets given in Theorem 5.1.

5.1. *Facets common to all hierarchical models.* Let $\mathcal{D}$ be the set of subsets of V defining the hierarchical model. Let $\mathcal{A}$ be the family of maximal elements of $\mathcal{D}$. For the subclass of graphical models Markov with respect to a graph G, $\mathcal{A}$ is the set of cliques of G. This set is traditionally denoted $\mathcal{C}$, but in this particular subsection, to avoid confusion between a clique $C \in \mathcal{C}$ and the polytope C, we use the notation $A \in \mathcal{A}$.

Let X be the design matrix given in Proposition 2.1 with rows equal to $(1, f_i^t)$, $i \in I$, and columns indexed by $J_0 = \varnothing \cup J$. Let X_{J_0} be the submatrix of X obtained by selecting the rows and columns of X indexed by J_0. Its inverse matrix $X_{J_0}^{-1}$ also has its rows and columns indexed by J_0. Let h_0, h_j, $j \in J$, denote the columns of $X_{J_0}^{-1}$. For any $j_0 \in J_0$ and $D \in \mathcal{D}$ such that $S(j_0) \subset D$, consider the vector $g_{j_0,D} \in \mathbb{R}^{J_0}$ defined as follows:

$$(g_{j_0,D})_j = \begin{cases} (h_{j_0})_j, & \text{if } j_0 \lhd j \text{ and } S(j) \subset D, \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

The vector $g_{j_0,D}$ is a subvector of h_{j_0} "padded" with zeros to obtain a vector in $\mathbb{R}^{J_0}$. Since $X_{J_0}^{-1}$ is given by the Moebius function of the partial order on J_0, the vectors (5.1) define the following linear forms in $\tilde{m} = (1, m_j, j \in J)$

$$\langle g_{0,D}, \tilde{m}\rangle = 1 + \sum_{j; S(j) \subset D} (-1)^{|S(j)|} m_j,$$

$$\langle g_{j_0,D}, \tilde{m}\rangle = \sum_{j; S(j) \subset D, j_0 \lhd j} (-1)^{|S(j)| - |S(j_0)|} m_j, \qquad j_0 \neq \varnothing.$$

We prefer to think of $\langle g_{0,D}, \tilde{m}\rangle$ and $\langle g_{j_0,D}, \tilde{m}\rangle$ as, respectively, affine and linear forms in $m = (m_j, j \in J)$, and we write

$$g_{0,D}(m) = 1 + \sum_{j;S(j)\subset D} (-1)^{|S(j)|} m_j, \tag{5.2}$$

$$g_{j_0,D}(m) = \sum_{j;S(j)\subset D, j_0 \lhd j} (-1)^{|S(j)|-|S(j_0)|} m_j, \qquad j_0 \neq \varnothing. \tag{5.3}$$

In this subsection, we will use $g_{j,A}$ only for $A \in \mathcal{A}$, but as we shall see in Section 5.2, $g_{j,S}$ when S is a minimal separator plays an important role also even though $S \notin \mathcal{A}$. In the next theorem, for $A \in \mathcal{A}$ and j such that $S(j) \subset A$, we consider the following affine hyperplanes of $\mathbb{R}^J$:

$$H(j, A) = \{m \in \mathbb{R}^J; g_{j,A}(m) = 0\}, \qquad j \in J \cup \{0\},$$

and we prove that

$$F(j, A) = H(j, A) \cap \overline{C} \tag{5.4}$$

is a facet of $\overline{C}$. Recall that for $T \subset V$, we use the notation $I_T = \prod_{v\in T} I_v$.

THEOREM 5.1. *Let A be in the set $\mathcal{A}$ of maximal elements of $\mathcal{D}$ defining a general hierarchical model. Let $j_0 \in J \cup \{0\}$ such that $S(j_0) \subset A$, and let $i \in I$. Then $g_{j_0,A}(f_i)$ can only take values* 0 *or* 1. *More precisely, the following holds*:

(1) *$g_{j_0,A}(f_i) = 1$ if and only if $j_0 \lhd i$ and $S(i) \cap A = S(j_0)$*;
(2) *there are exactly $|I| - |I_{V\setminus A}|$ vectors f_i's such that $g_{j_0,A}(f_i) = 0$*;
(3) *the set $F(j_0, A)$ as defined in* (5.4) *is a facet of the polytope $\overline{C}$*.

The proof of the theorem is given in the supplementary file [15]. The proof of parts (1) and (2) is straightforward and follow from the Moebius form of the equation of the facets. The proof of part (3) is long and technical, but its idea is very simple: we know from parts (1) and (2) that $\overline{C}$ is supported by $H(j_0, A)$; we then show that if $h \in H(j_0, A)$ is orthogonal to all the f_i contained in $F(j_0, A)$, then $h = 0$, and therefore these f_i's affinely generate $H(j_0, A)$, and $F(j_0, A)$ is a facet of $\overline{C}$.

Let us illustrate the results in Theorem 5.1 in the following example. We consider the model studied in Example 2.1 and list the various faces and the f_i's that belong to them. The two vertical and horizontal lines in the two arrays below are only there for visual comfort.

EXAMPLE 5.1. The matrix X_{J_0} is therefore given by the following array:

	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

The matrix $X_{J_0}^{-1}$ is given by:

	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 two maximal elements in $\mathcal{A}$ are ab and bc, and

$$m = (m_{100}, m_{200}, m_{010}, m_{020}, m_{110}, m_{210}, m_{120}, m_{220}, m_{001}, m_{011}, m_{021}).$$

The equation of the facets with $A = ab$ are obtained by following the definition (5.1) of $g_{j_0,A}$:

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

$$g_{210,ab} = m_{210},$$
$$g_{120,ab} = m_{120},$$
$$g_{220,ab} = m_{220}.$$

The equation of the facets with $A = bc$ follows a similar pattern, and as we shall see in Corollary 5.1, these are the only facets of C.

Let $F_{j_0,A}$ denote the facet given by $\overline{C} \cap H_{j_0,A}$. The facets can also be described by their extreme points f_i. It is easier to give those f_i *not* in the facet $F_{j_0,A}$. For $F_{\varnothing,ab}$, f_0 and f_{001} are the only f_i not in the face. For $F_{100,ab}$, f_{100} and f_{101} are the only f_i not in the face while for $F_{120,ab}$, f_{120} and f_{121} are the absent vectors.

5.2. *Facets of $\overline{C}$ when G is decomposable*. When the graph G is decomposable, the normalizing constant $I(m,\alpha)$ is the normalizing constant of the hyper Dirichlet as defined in [5]. In the theorem below, we restate, in our present notation, the expression of $I(m,\alpha)$ as given in formula (4.8) of [16] and directly derive the form of $\mathbb{J}_C(m)$ for decomposable models. A corollary giving the facets of $\overline{C}$ when the model is decomposable follows immediately from the theorem.

THEOREM 5.2. *Let $(V,\mathcal{E})$ be a decomposable graph, let $\mathcal{C}$ be the family of its cliques, let $\mathcal{S}$ be the family of its minimal separators and let $\nu(S)$ be the multiplicity of the minimal separator S. Then for m in the interior C of the convex hull of the f_i's, we have*

$$\begin{aligned} I(m,\alpha) &= \int_{\mathbb{R}^J} e^{\alpha\langle\theta,m\rangle} L(\theta)^{-\alpha}\, d\theta \\ &= \frac{\prod_{C\in\mathcal{C}} \Gamma(\alpha g_{0,C}(m)) \prod_{\{j\in J; S(j)\subset C\}} \Gamma(\alpha g_{j,C}(m))}{\Gamma(\alpha)\prod_{S\in\mathcal{S}}[\Gamma(\alpha g_{0,S}(m))\prod_{\{j\in J;S(j)\subset S\}}\Gamma(\alpha g_{j,S}(m))]^{\nu(S)}} \end{aligned} \tag{5.5}$$

and

$$\begin{aligned} \lim_{\alpha\to} \alpha^{|J|} I(m,\alpha) &= \mathbb{J}_C(m) \\ &= \frac{\prod_{S\in\mathcal{S}}[g_{0,S}(m)\prod_{\{j\in J;S(j)\subset S\}} g_{j,S}(m)]^{\nu(S)}}{\prod_{C\in\mathcal{C}} g_{0,C}(m)\prod_{\{j\in J;S(j)\subset C\}} g_{j,C}(m)}. \end{aligned} \tag{5.6}$$

COROLLARY 5.1. *In the case of a hierarchical model associated to a decomposable graph, all the facets of $\overline{C}$ are of the type $F(j_0, C)$ described in Theorem 5.1, with $j_0 \in J$, with C in the set $\mathcal{C}$ of cliques and $S(j_0) \subset C$.*

PROOF. We know from Theorem 5.1 that the affine forms in the denominator of $\mathbb{J}_C(m)$ in (5.6) define facets of $\overline{C}$. From Theorem 3.1, we know that they are the only ones. $\square$

In fact we conjecture, as mentioned in the Introduction, that if a model is such that the only facets of $\overline{C}$ are of the type given in Theorem 5.1, then it is a decomposable graphical model.

EXAMPLE 5.2. If $V = \overset{a}{\bullet} - \overset{b}{\bullet} - \overset{c}{\bullet}$ and if $I = \{0,1,2\} \times \{0,1\} \times \{0,1\}$, we have

$$\begin{aligned}
g_{0,bc}(m) &= 1 - m_{001} - m_{010} + m_{011},\\
g_{001,bc}(m) &= m_{001} - m_{011},\\
g_{010,bc}(m) &= m_{010} - m_{011},\\
g_{011,bc}(m) &= m_{011},\\
g_{0,ab}(m) &= 1 - m_{100} - m_{200} - m_{010} + m_{110} + m_{210},\\
g_{100,ab}(m) &= m_{100} - m_{110},\\
g_{200,ab}(m) &= m_{200} - m_{210},\\
g_{010,ab}(m) &= m_{010} - m_{110} - m_{210},\\
g_{110,ab}(m) &= m_{110},\\
g_{210,ab}(m) &= m_{210},\\
g_{0,b}(m) &= 1 - m_{010},\\
g_{010,b}(m) &= m_{010}.
\end{aligned}$$

In this case $I(m,\alpha)$ is a quotient: the numerator is the product of 10 gamma functions and the denominator is $\Gamma(\alpha)\Gamma(\alpha(1-m_{010}))\Gamma(\alpha m_{010})$. As a consequence, $\mathbb{J}_C(m)$ is

$$\begin{aligned}
&(g_{0,b}(m) g_{010,b}(m))\\
&\quad\times \big(g_{0,bc}(m) g_{001,bc}(m) g_{010,bc}(m) g_{011,bc}(m) g_{0,ab}(m) g_{100,ab}(m)\\
&\qquad\qquad \times g_{200,ab}(m) g_{010,ab}(m) g_{110,ab}(m) g_{210,ab}(m)\big)^{-1}.
\end{aligned}$$

5.3. *Facets of $\overline{C}$ when the model is binary and the model is governed by a cycle.* For the sake of completion and the convenience of the reader, we recall some known results giving the facets of the polytope $\overline{C}$ when the model is hierarchical, binary and governed by a cycle G of order $n \geq 3$. The reader is referred to Theorem 27.3.3 in [6] and [13] and some references within for an explicit description of these facets. In this subsection, we will simply translate the equation of the facets given there in our own coordinates. The results are given in the following theorem. The coordinates of $m \in \mathbb{R}^J$ will be denoted m_v if they are indexed by a vertex $v \in V$ and by m_e if they are indexed by an edge $e \in E$.

THEOREM 5.3. *Let $G = (V, E)$ be a cycle of order $n \geq 3$. Assume the hierarchical model is binary and governed by G, that is, $\mathcal{D} = \{v \in V, e \in E\}$. Then the polytope $\overline{C}$ is defined by the following equations and the facets are defined by the corresponding equalities*:

(1) *for any edge $(a, b) \in E$,*

$$m_{ab} \geq 0, \qquad m_a - m_{ab} \geq 0, \tag{5.7}$$

$$m_b - m_{ab} \geq 0, \qquad 1 - m_a - m_b + m_{ab} \geq 0, \tag{5.8}$$

(2) *for any subset $F \subseteq E$ with odd cardinality $|F|$,*

$$\sum_{(a,b)\in F} (m_a + m_b - 2m_{ab}) - \left(\sum_{v\in V} m_v - \sum_{e\in E} m_e\right) \leq \frac{|F| - 1}{2}. \tag{5.9}$$

The total number of facets for the polytope $\overline{C}$ of the model governed by the cycle of order n is $F_n = \sum_{k\in N, k \text{ odd}, k\leq n} \binom{n}{k}$.

We see that the facets given by the first four equations are those described in Theorem 5.1 corresponding to the cliques $\{(a, b) \in E\}$ while the others are specific to models governed by a cycle. We illustrate this theorem in the case of the cycles of order 3, 4 and 5. We will not repeat the facets (5.7) and (5.8) common to all hierarchical models. We will give the facets of type (5.9) only.

For $n = 3$, let $V = \{a, b, c\}$ and $E = \{(a, b), (b, c), (c, a)\}$. The four facets of type (5.9) are

$$\begin{aligned} 1 - m_a - m_b - m_c + m_{ab} + m_{bc} + m_{ac} &\geq 0, \\ m_{ab} + m_c - m_{bc} - m_{ac} &\geq 0, \end{aligned} \tag{5.10}$$

and the other two facets obtained from (5.10) by permutations of the edges of G.

For $n = 4$, let $V = \{a, b, c, d\}$ and $E = \{(a, b), (b, c), (c, d), (d, a)\}$. The eight facets of type (5.9) are

$$1 - m_b - m_c + m_{ab} + m_{bc} + m_{cd} - m_{da} \geq 0, \tag{5.11}$$

$$m_c + m_d + m_{ab} - m_{bc} - m_{cd} - m_{da} \geq 0, \tag{5.12}$$

and the other three facets obtained from each of (5.11) and (5.12) by permutations of the edges of G.

For $n = 5$, let $V = \{a, b, c, d, e\}$ and $E = \{(a, b), (b, c), (c, d), (d, e), (e, a)\}$. The sixteen facets of type (5.9) are

$$m_{ab} + m_c + m_d + m_e - m_{bc} - m_{cd} - m_{de} - m_{da} \geq 0, \tag{5.13}$$

$$1 - m_a - m_b + m_{ea} + m_{ab} + m_{bc} + m_d - m_{cd} - m_{ed} \geq 0, \tag{5.14}$$

$$1 - m_d + m_{ab} + m_{cd} + m_{de} - m_{bc} - m_{ae} \geq 0, \tag{5.15}$$

$$2 - m_a - m_b - m_c - m_d - m_e + m_{ab} + m_{bc} + m_{cd} + m_{de} + m_{ea} \geq 0,$$

and the other four facets obtained from each of (5.13), (5.14) and (5.15) by permutations of the edges of G.

6. Conclusion. Our paper gives the description of the behavior of the Bayes factor as $\alpha \to 0$. We have shown that, in this study, what counts is the dimension of the face to which the data belongs rather than the dimension of the model. Moreover, it is not surprising to see that $\overline{C}$, the convex hull of the support of the generating measure of the multinomial for the hierarchical model, is important since the multinomial is a natural exponential family. We know that it is equally important in the study of the existence of the maximum likelihood estimate of the parameter; see, for example, Eriksson et al. [9], Geiger et al. [11] or Rinaldo [18]. However, the role of the characteristic function $\mathbb{J}_C(\cdot)$ of C has only been uncovered here in the study of the Bayes factor, and we can add $\mathbb{J}_C$ to the toolkit of exponential families. It is remarkable that in our case, when $\overline{C}$ is a bounded polytope, $\mathbb{J}_C$ can be expressed as a rational function such that its denominator describes the facets of $\overline{C}$.

We note that Theorem 3.1 is proved under the assumption that the polytope C is bounded, and our present result are valid for the multinomial only, but we believe that they can be extended to the case when the sampling distribution is Poisson (and also product multinomial). This is the topic of current work.

A secondary contribution of this paper is our results on the identification of the facets of a polytope. We have two new results for polychotomous models (i.e., not necessarily binary): the first giving a particular category of facets common to all hierarchical models, the second giving the complete set of facets for decomposable models.

For decomposable models, we extend the notion of effective degree of freedom given in [20] and give a quick way to predict the behavior of the Bayes factor without using the concept of face or facets of a polytope.

APPENDIX: PROOF OF THEOREM 3.3

Without loss of generality, we assume that $m = 0$ so that $\mathbb{J}_C(\lambda m + (1-\lambda)y) = \mathbb{J}_C((1-\lambda)y)$. From (3.3) we have

$$\text{(A.1)} \qquad \frac{J_C((1-\lambda)y)}{n!} = \int_{C^o} \frac{d\theta}{(1-(1-\lambda)\langle\theta, y\rangle)^{n+1}}.$$

Recall that C^0 is closed. In order to study the behavior of this last integral when $\lambda \to 0$, we are going to build a parametrization of C^o which gives a special role to $\widehat{F}$, the face of C^o dual to the face F of $\overline{C}$ containing y in its interior.

Let $\mathcal{E}$ the set of extreme points of $\overline{C}$ and $\mathcal{I} \subset \mathcal{E}$ the set of extreme points of F. To F we associate the dual face of C^o defined by

$$\text{(A.2)} \qquad \widehat{F} = \{\theta \in C^o | \langle\theta, f\rangle = 1\ \forall f \in \mathcal{I}\}.$$

It is a classical result (see [3]) that $\widehat{F}$ has dimension $n - k - 1$. Let us now observe that we have an equivalent representation of $\widehat{F}$ in (A.2) as

$$\text{(A.3)} \qquad \widehat{F} = \{\theta \in C^o | \langle\theta, y\rangle = 1\}.$$

Indeed, since y is in the relative interior of F we write

$$y = \sum_{f \in \mathcal{I}} \lambda_f f,$$

where $\lambda_f > 0$ and $\sum_{f \in \mathcal{I}} \lambda_f = 1$. Here $\lambda_f > 0$ is important in the argument to follow. Clearly $\widehat{F} \subset \{\theta \in C^o; \langle \theta, y \rangle = 1\}$. Conversely if $\langle \theta, y \rangle = 1$ then $\sum_{f \in \mathcal{I}} \lambda_f (1 - \langle \theta, f \rangle) = 0$. If furthermore $\theta \in C^o$ we have $1 - \langle \theta, f \rangle \geq 0$ and therefore $1 - \langle \theta, f \rangle = 0$ which shows $\widehat{F} \supset \{\theta \in C^o; \langle \theta, y \rangle = 1\}$ and proves (A.3).

Next, for $\varepsilon > 0$ small, we consider the following approximation $\widehat{F}_\varepsilon$ of $\widehat{F}$

$$\widehat{F}_\varepsilon = \{\theta \in C^o; \langle \theta, y \rangle = 1 - \varepsilon\}, \tag{A.4}$$

which is a $(n-1)$-dimensional convex subset of C^o and we want to prove that $\mathrm{vol}_{n-1} \widehat{F}_\varepsilon \sim c\varepsilon^k$ for some positive constant c. Using (A.3) and(A.2), we can rewrite (A.4) as

$$\widehat{F}_\varepsilon = \left\{\theta \in C^o; \sum_{f \in \mathcal{I}} \lambda_f (1 - \langle \theta, f \rangle) = \varepsilon\right\}. \tag{A.5}$$

To show $\mathrm{vol}_{n-1} \widehat{F}_\varepsilon \sim c\varepsilon^k$ we parametrize $\widehat{F}_\varepsilon$ as follows: let $\theta \mapsto x = \varphi(\theta)$ be the affine map from E^* to $\mathbb{R}^{\mathcal{I}}$ defined by

$$x_f = \lambda_f (1 - \langle \theta, f \rangle), \qquad f \in \mathcal{I}, \tag{A.6}$$

which is equivalent to $\langle \theta, f \rangle = 1 - \frac{x_f}{\lambda_f}$. The set $S_\varepsilon = \varphi(\widehat{F}_\varepsilon)$ is therefore the intersection of the simplex

$$\left\{x \in \mathbb{R}^{\mathcal{I}}; x_f \geq 0 \ \forall f \in \mathcal{I}, \sum_{f \in \mathcal{I}} x_f = \varepsilon\right\} \tag{A.7}$$

and of the convex set $\varphi(C^o)$ which is contained in the affine manifold $\varphi(E^*) \subset \mathbb{R}^{\mathcal{I}}$. If $x \in S_\varepsilon$ then its preimage by φ is the set

$$\varphi^{-1}(x) = \left\{\theta \in E^*; \langle \theta, f \rangle = 1 - \frac{x_f}{\lambda_f} \ \forall f \in \mathcal{I}\right\},$$

which is an affine subspace of E^* parallel to the linear space

$$H_0 = \{\theta \in E^*; \langle \theta, f \rangle = 0 \ \forall f \in \mathcal{I}\}, \tag{A.8}$$

which has dimension $n - k - 1$ since F has dimension k. As a result we can write $\widehat{F}_\varepsilon$ as the following union of disjoint sets

$$\widehat{F}_\varepsilon = \bigcup_{x \in S_\varepsilon} (\varphi^{-1}(x) \cap C^o), \tag{A.9}$$

which is saying that $\widehat{F}_\varepsilon$ can be parametrized by (x, z) where $x \in S_\varepsilon$, a convex set of dimension k, and where $z \in \varphi^{-1}(x) \cap C^o$, a convex set of dimension $n - k - 1$.

The bijection $\theta \mapsto (x, z)$ is the restriction to $\widehat{F}_\varepsilon$ of an affine map and therefore its Jacobian K such that $d\theta = K\,dx\,dz$ is a constant.

$$\mathrm{vol}_{n-1}\widehat{F}_\varepsilon = \int_{\widehat{F}_\varepsilon} d\theta = K \int_{S_\varepsilon} \left(\int_{\varphi^{-1}(x)\cap C^o} dz \right) dx.$$

If we fix x^0 in the simplex (A.7), then the behavior of $\int_{\varphi^{-1}(\varepsilon x^0)\cap C^o} dz$ is easy to describe since $\lim_\varepsilon \varphi^{-1}(\varepsilon x^0) \cap C^o = \widehat{F}$ in the sense of polytopes, which implies

$$\lim_{\varepsilon\to 0} \int_{\varphi^{-1}(\varepsilon x^0)\cap C^o} dz = \lim_{\varepsilon\to 0} \mathrm{vol}_{n-k-1}\big(\varphi^{-1}(\varepsilon x^0) \cap C^o\big) = \mathrm{vol}_{n-k-1}(\widehat{F}).$$

Let us now observe that 0 is an extreme point of $\varphi(C^o)$. If not there exist $x = \varphi(\theta)$ and $x' = \varphi(\theta')$ with θ and $\theta' \in C^o$ such that $x + x' = 0$, that is, for all $f \in \mathcal{I}$

$$1 - \lambda_f\langle\theta, f\rangle + 1 - \lambda_f\langle\theta', f\rangle = 2 - \lambda_f[\langle\theta, f\rangle + \langle\theta', f\rangle] = 0.$$

Since $0 \le \lambda_f \le 1$, this in turn implies $\lambda_f = 1$ and $\langle\theta, f\rangle + \langle\theta', f\rangle = 2$. Since $\langle\theta, f\rangle$ and $\langle\theta', f\rangle$ are ≤ 1 this implies $x_f = x'_f = 0$ for all $f \in \mathcal{I}$, a contradiction. Now we use the fact that C^o is a polytope and so is $\varphi(C^o)$ which has dimension $k+1$. For ε small enough (say $0 < \varepsilon \le \varepsilon_0$) the intersection S_ε of the simplex given in (A.7) with $\varphi(C^o)$ coincides with the intersection of the simplex with the support cone of $\varphi(C^o)$ at its vertex 0. Since a cone is invariant by dilations we can claim that there exists a number $c_1 > 0$ such that for $0 < \varepsilon \le \varepsilon_0$ we have $\mathrm{vol}_k(S_\varepsilon) = c_1\varepsilon^k$. Finally

$$\mathrm{vol}_{n-1}\widehat{F}_\varepsilon \sim c_1 K\,\mathrm{vol}_{n-k-1}(\widehat{F})\varepsilon^k. \tag{A.10}$$

The parametrization of θ in (A.1) is therefore (x, z, ε) where (x, z) is as given in (A.9) and the range of ε is such that, for that range, F_ε describes all of C^o. We note that the bounded function $\mathrm{vol}_{n-1}\widehat{F}_\varepsilon = f(\varepsilon)$ is zero if ε is big enough since then $\widehat{F}_\varepsilon$ becomes empty and, of course, $\mathrm{vol}_{n-1}\widehat{F}_0 = \mathrm{vol}_{n-1}\widehat{F}$. Let b be such that $f(\varepsilon) = 0$ when $\varepsilon > b$. When ε varies from 0 to $+\infty$, $\widehat{F}_\varepsilon$ generates all of C^o. Then, following (A.10), equation (A.1) becomes

$$\begin{aligned}\int_{C^o} \frac{d\theta}{(1-(1-\lambda)\langle\theta, y\rangle)^{n+1}} &= \int_0^\infty \frac{\mathrm{vol}_{n-1}\widehat{F}_\varepsilon\,d\varepsilon}{(1-(1-\lambda)(1-\varepsilon))^{n+1}} \\ &= \int_0^\infty \frac{f(\varepsilon)\,d\varepsilon}{(1-(1-\lambda)(1-\varepsilon))^{n+1}}.\end{aligned}$$

Using $f(\varepsilon) \sim c\varepsilon^k$ we will now show that

$$\lim_{\lambda\to 0} \lambda^{n-k} \int_0^\infty \frac{f(\varepsilon)\,d\varepsilon}{(1-(1-\lambda)(1-\varepsilon))^{n+1}} = cB(k+1, n-k), \tag{A.11}$$

which concludes the proof. To derive (A.11), we first show that for $0 < a < b$:

(1) $\lambda^{n-k}\int_0^a \frac{\varepsilon^k\,d\varepsilon}{(\lambda+\varepsilon-\lambda\varepsilon)^{n+1}} \to_{\lambda\to 0} B(k+1, n-k)$,

(2) $\lim_{\lambda\to 0}\lambda^{n-k}\int_a^b \frac{d\varepsilon}{(\lambda+\varepsilon-\lambda\varepsilon)^{n+1}} = 0$.

AOS imspdf v.2012/03/16 Prn:2012/03/29; 11:11 F:aos974.tex; (Laima) p. 29

Statement (1) is shown by the change of variable $\varepsilon = \lambda t$ and the theorem of dominated convergence. Indeed, for $0 < \lambda \le \lambda_0 < 1$, we have

$$\lambda^{n-k} \int_0^a \frac{\varepsilon^k \, d\varepsilon}{(\lambda + \varepsilon - \lambda\varepsilon)^{n+1}} = \int_0^{a/\lambda} \frac{t^k dt}{(1 + t - \lambda t)^{n+1}} \le \int_0^{a/\lambda} \frac{t^k dt}{(1 + t - \lambda_0 t)^{n+1}},$$

which tends to $\frac{1}{(1-\lambda_0)^{k+1}} B(k+1, n-k)$ when $\lambda \to 0$. Since this is true for any $\lambda_0 > 0$, statement (1) follows.

Statement (2) is obvious since $\int_a^b \frac{d\varepsilon}{(\lambda+\varepsilon-\lambda\varepsilon)^{n+1}} < \int_a^b \frac{d\varepsilon}{\varepsilon^{n+1}}$ is finite.

Next, fix $\delta > 0$. There exists $a < b$ such that $|\frac{f(\varepsilon)}{\varepsilon^k} - c| \le \delta$ if $0 < \varepsilon \le a$. Writing this as $-\delta\varepsilon^k < f(\varepsilon) - c\varepsilon^k < \delta\varepsilon^k$, integrating and using (1) yields

$$\limsup_{\lambda \to 0} \left| \frac{1}{B(k+1, n-k)} \int_0^a \frac{f(\varepsilon)\, d\varepsilon}{(1 - (1-\lambda)(1-\varepsilon))^{n+1}} - c \right| \le \delta.$$

Since f is bounded (2) implies that

$$\begin{aligned} &\limsup_{\lambda \to 0} \lambda^{n-k} \int_a^{+\infty} \frac{f(\varepsilon)\, d\varepsilon}{(1 - (1-\lambda)(1-\varepsilon))^{n+1}} \\ &\quad = \limsup_{\lambda \to 0} \lambda^{n-k} \int_a^{b} \frac{f(\varepsilon)\, d\varepsilon}{(1 - (1-\lambda)(1-\varepsilon))^{n+1}} = 0. \end{aligned}$$

Thus for all $\delta > 0$ we have

$$\limsup_{\lambda \to 0} \left| \frac{1}{B(k+1, n-k)} \int_0^\infty \frac{f(\varepsilon)\, d\varepsilon}{(1 - (1-\lambda)(1-\varepsilon))^{n+1}} - c \right| \le \delta,$$

which implies (A.11).

Acknowledgments. We are grateful to Steffen Lauritzen who asked the question which motivated this paper, that is, what happens to the Bayes factor when $\alpha \to 0$, and who told us about [20]. We thank Seth Sullivant for pointing out the work of [6] and [13], Jean-Baptiste Hiriart-Urruty for refering us to [1], Alexander Barvinok who directed us to the second formula of (3.3), Mathieu Meyer who suggested a proof of Theorem 3.2 much shorter than our initial one, Monique Laurent who helped us with Theorem 5.3 and Johannes Rauh who pointed out several typographical errors, minor inaccuracies and suggested a few shortcuts in some proofs. Last but not least, we thank the Referee, the Associate Editor and the Editor for reading our paper so carefully. We are particularly grateful to the Referee who did a masterful job of reading and understanding our paper, who suggested a better presentation and helped us highlight some important ideas.

SUPPLEMENTARY MATERIAL

Proofs (DOI: 10.1214/12-AOS974SUPP; .pdf). This section contains a characterization of the hierarchical loglinear model as well as the statement and proofs of Lemmas 3.1, 3.3 and Theorems 3.1, 4.1 and 5.1. aos974_supp.pdf

REFERENCES

[1] AZÉ, D. and HIRIART-URRUTY, J. B. (1994). *Analyse Variationnelle et Optimisation*. Cépaduès, Toulouse. <auto>

[2] BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333 <mr>

[3] BARVINOK, A. (2002). *A Course in Convexity*. *Graduate Studies in Mathematics* **54**. Amer. Math. Soc., Providence, RI. MR1940576 <mr>

[4] DARROCH, J. N. and SPEED, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11** 724–738. MR0707924 <mr>

[5] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 <mr>

[6] DEZA, M. M. and LAURENT, M. (1995). *Geometry of Cuts and Metrics*. Springer, New York. <auto>

[7] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238 <mr>

[8] EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. MR0801773 <mr>

[9] ERIKSSON, N., FIENBERG, S. E., RINALDO, A. and SULLIVANT, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.* **41** 222–233. MR2197157 <mr>

[10] FARAUT, J. and KORÁNYI, A. (1994). *Analysis on Symmetric Cones*. Oxford Univ. Press, New York. MR1446489 <mr>

[11] GEIGER, D., MEEK, C. and STURMFELS, B. (2006). On the toric algebra of graphical models. *Ann. Statist.* **34** 1463–1492. MR2278364 <mr>

[12] HAUGHTON, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16** 342–355. MR0924875 <mr>

[13] HOŞTEN, S. and SULLIVANT, S. (2002). Gröbner bases and polyhedral geometry of reducible and cyclic models. *J. Combin. Theory Ser. A* **100** 277–301. MR1940337 <mr>

[14] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991 <mr>

[15] LETAC, G. and MASSAM, H. (2012). Supplement to "Bayes factors and the geometry of discrete hierarchical loglinear models." DOI:10.1214/12-AOS974SUPP. <unstr>

[16] MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical loglinear models. *Ann. Statist.* **37** 3431–3467. MR2549565 <mr>

<uncited> [17] RAINVILLE, E. D. (1960). *Special Functions*. Macmillan Co., New York. MR0107725 <mr>

[18] RINALDO, A. (2006). On maximum likelihood estimation for log-linear models. Technical Report 833, Dept. Statistics, Carnegie Mellon Univ. <unstr>

[19] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014 <mr>

[20] STECK, S. and JAAKKOLA, T. S. (2002). On the Dirichlet prior and Bayesian regularization. MIT Memo. <unstr>

LABORATOIRE DE STATISTIQUE
ET PROBABILITÉ
UNIVERSITÉ PAUL SABATIER
TOULOUSE, 31000
FRANCE
E-MAIL: letac@cict.fr

DEPARTMENT OF MATHEMATICS
AND STATISTICS
YORK UNIVERSITY
TORONTO, M3J1P3
CANADA
E-MAIL: massamh@yorku.ca

THE ORIGINAL REFERENCE LIST

The list of entries below corresponds to the original Reference section of your article. The bibliography section on previous page was retrieved from MathSciNet applying an automated procedure.
Please check both lists and indicate those entries which lead to mistaken sources in automatically generated Reference list.

[1] AZÉ, D. AND HIRIART-URRUTY, J.-B. (1994). *Analyse variationnelle et optimisation,* Cépaduès, Toulouse.

[2] BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

[3] BARVINOK, A. (2002). *A Course on Convexity*, American Mathematical Society, Providence, Rhode Island.

[4] DARROCH, J. N. AND SPEED, T. P. (1983). Additive and multiplicative models and interactions, *Ann. Statist.*, **11**, 724–738.

[5] DAWID, A. P. AND LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.*, **21**, 1272 -1317.

[6] DEZA, M.-M. AND LAURENT, M. (1995). *Geometry of Cuts and Metrics*, Springer, New York.

[7] DIACONIS, P. AND YLVISAKER, D. (1979). Conjugate priors for exponential families, *Ann. Statist.*, **7**, 269 –281.

[8] EDWARDS, D. E. AND HAVRANEK, T. (1985). A fast procedure for model search in multidimensional contingency tables, *Biometrika*, **72**, 339 –351.

[9] ERIKSSON, N., FIENBERG, S.E., RINALDO, A. AND SULLIVANT, S., (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models, *Journal of Symbolic Computation*, **41**, 222 -233.

[10] FARAUT, J. AND KORANYI, A. (1994). *Analysis on Symmetric Cones,* Clarendon Press, Oxford.

[11] GEIGER, A., MEEK, C. AND STURMFELS, B. (2006). On the toric algebra of graphical models, *Ann. Statist.*, **34**, 1463 - 1492.

[12] HAUGHTON, D.M.A., (1988). On the Choice of a Model to Fit Data from an Exponential Family, *Ann. Statist.*, **16**, 342 -355.

[13] HOSTEN, S. AND SULLIVANT, S. (2002). Grobner bases and polyhedral geometry of reducible and cyclic models, *J. Combinatorial Theory*,**100**, 277-301.

[14] LAURITZEN, S.L. (1996). *Graphical Models*, Oxford Science Publications.

[15] LETAC, G. AND MASSAM, H. (2012). Supplement to “Bayes factors and the geometry of discrete hierarchical loglinear models.” DOI:???

[16] MASSAM, H., LIU, J. AND DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models, *Ann. Statist.*, **37**, 3431 - 3467.

[17] RAINVILLE, E. D. (1960). *Special Functions*, Macmillan, New York.

[18] RINALDO, A. (2006). On maximum likelihood estimation for log-linear models, *Technical Report 833, Department of Statistics, Carnegie Mellon University*.

[19] SCHWARZ, G., Estimating the dimension of the model, *Ann. Statist.*, **6**, 461-464.

[20] STECK, S. AND JAAKKOLA, T.S. (2002). On the Dirichlet prior and Bayesian regularization, *MIT Memo*.

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Bayes factors and the geometry of discrete hierarchical log-linear models
Author : G\351rard Letac, H\351l\350ne Massam
Subject : The Annals of Statistics, 0, Vol.0, No.00, 1-30
Keywords: 62H17, 62C10, 62J12, Bayes, sparse contingency tables, conjugate priors, characteristic function of a convex set, Dirichlet distribution

THE LIST OF URI ADRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page:
http://www.e-publications.org/ims/support/urihelp.html.

```
200 http://www.imstat.org/aos/ [2:pp.1,1] OK
404 http://dx.doi.org/10.1214/12-AOS974 [2:pp.1,1] Not Found
200 http://www.imstat.org [2:pp.1,1] OK
200 http://www.ams.org/mathscinet/msc/msc2010.html [2:pp.1,1] OK
404 http://dx.doi.org/10.1214/12-AOS974SUPP [4:pp.30,30,32,32] Not Found
--- mailto:letac@cict.fr [2:pp.33,33] Check skip
--- mailto:massamh@yorku.ca [2:pp.33,33] Check skip
```