Bayes Factors for Goodness of Fit Testing

Fulvio Spezzaferri, Isabella Verdinelli, and Massimo Zeppieri
Università di Roma *La Sapienza* and Carnegie Mellon University

October 31, 2003

We propose the use of the generalized fractional Bayes factor for testing fit in multinomial models. This is a non-asymptotic method that can be used to quantify the evidence for or against a sub-model. We give expressions for the generalized fractional Bayes factor and we study its properties. In particular, we show that the generalized fractional Bayes factor has better properties than the fractional Bayes factor.

Keywords: generalized fractional Bayes factor, Dirichlet process, Beta-Stacy process.

1. Introduction.

In this paper we propose a Bayesian method for testing fit in multinomial models. Specifically, we will use the Bayes factor for evaluating the evidence for or against a null sub-model of the multinomial.

The advantages of using a Bayesian approach for this problem are that it does not rely on large sample asymptotics and it provides a measure of evidence in favor of the null model in contrast to p-values which are usually only regarded as measures of evidence against the null. The main disadvantage of the Bayesian method is that it requires specifying a prior distribution. When a defensible, subjective prior is available, the Bayesian approach is straightforward: we simply compute the Bayes factor, the posterior odds in favor of the null model divided by the prior odds. When a subjective prior is not available, one possibility is to use some sort of objective Bayesian procedure. These include the work on intrinsic Bayes factor, first introduced by Berger and Pericchi (1996a, b), and also examined by other authors (for example O′Hagan, 1997; Moreno, Bertolino and Racugno,1998), the fractional Bayes factor (O′Hagan 1991, 1995, 1997), and the generalized fractional Bayes factor (De Santis and Spezzaferri, 1997, 2001).

The intrinsic Bayes factor has been shown to have good properties but it is computationally intensive and in the case of discrete data the presence of empty cells might pose some difficulties. We thus focus on the fractional Bayes factor and the generalized fractional Bayes factor.

The main goals of this paper are (i) to give closed form expressions for the generalized fractional Bayes factor and (ii) to show that the generalized fractional Bayes factor has better statistical behavior than the fractional Bayes factor.

Goodness of fit testing in the Bayesian framework for continuous data has been examined by a number of authors (for example Verdinelli and Wasserman, 1998; Berger and Guglielmi, 2001; Robert and Rousseau, 2003). For discrete data, the paper by Albert (1997) illustrates

procedures for testing and estimating association structures in contingency tables. A review of Bayesian and non-Bayesian alternatives to the χ^2 method is presented in Conigliani, Castro and O'Hagan (2000). Our work concentrates on discrete data and, to a large degree, builds on and improves on Conigliani, Castro and O'Hagan (CCO; 2000). The present paper also highlights ways for using hierarchical prior distribution based on the Dirichlet process as well as on the Beta-Stacy process considered in Walker and Muliere (1997)

In Section 2 we briefly illustrate the CCO procedure for goodness of fit testing, based on the fractional Bayes factor and on using a Dirichlet distribution for modeling a two-stage hierarchical prior on the alternative model. We also introduce the generalized fractional Bayes factor and present examples and general results showing that the generalized fractional Bayes factor has good statistical properties.

In Section 3 we consider a hierarchical prior based on the Dirichlet model, and obtain an explicit expression for the generalized fractional Bayes factor. Examples are also considered for comparing the two methods presented in the paper.

In Section 4, an extension to using a Beta-Stacy process (Walker and Muliere, 1997) for modeling the hierarchical prior is considered. Section 5 contains a brief discussion.

2. Goodness of fit testing: Bayesian alternatives to the χ^2.

2.1 Introduction.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a sample of n iid observations classified in one of $k+1$ groups, denoted by G_j $(j = 0, \cdots, k)$, and let $\mathbf{r} = (r_0, \ldots, r_k)$ be a vector whose entries are the number of x_i's falling in each group.

Conigliani, Castro and O'Hagan (2000) presented the following Bayesian approach to goodness of fit testing. Let $\mathrm{MN}(\boldsymbol{\alpha})$ denote a multinomial distribution with parameters $\boldsymbol{\alpha} = (\alpha_0, \cdots \alpha_k)$, $\sum \alpha_j = 1$ and let $Pr(x_s \in G_j) = \alpha_j, s = 1, \cdots, n$, with $\mathbf{r} \sim \mathrm{MN}(\boldsymbol{\alpha})$). They assumed that under M_1 the α_j's belong to a specific parametric family $\mathcal{F} = \{\alpha_j(\phi), \phi \in \mathbb{R}\}$, characterized by $\alpha_j(\phi)$ depending on a parameter ϕ, while under M_2, no assumptions on the form of α_j's were made. The alternative model M_2 is thus equivalent to the alternative hypothesis implicitely used in the χ^2 method, where data can be generated by any probability distribution. A Bayesian hypothesis testing procedure for choosing between models M_1 and M_2 will give a Bayesian non-parametric version of the χ^2 goodness of fit test.

The two models that CCO compared are:

$$M_1 : \mathbb{P}(x_s \in G_j|\phi) = \alpha_j(\phi) \qquad M_2 : \mathbb{P}(x_s \in G_j|\boldsymbol{\alpha}) = \alpha_j \tag{1}$$

The parameters $\alpha_j(\phi)$ in M_1 depend on ϕ, and CCO considered ϕ to be distributed according to a noninformative prior $\pi(\phi)$. They also assumed that, under M_2, the vector $\boldsymbol{\alpha}$ had a Dirichlet prior distribution with parameters $(c_0, \cdots, c_k)$ and density:

$$\pi_2(\boldsymbol{\alpha}|c) = \frac{\Gamma(c)}{\prod_{j=0}^{k} \Gamma(c_j)} \prod_{j=0}^{k} \alpha_j^{c_j - 1}, \tag{2}$$

where $c = \sum c_j$, $(c_j > 0; j = 0, \cdots, k)$. The assumption $\mathbb{E}[\alpha_j] = \alpha_j(\phi)$ was also made to express the idea that the alternative model M_2 was *centered* on the null model M_1, so that, from $\mathbb{E}[\alpha_j] = c_j/c$, it follows $c_j = c\alpha_j(\phi)$. As they did for model M_1, CCO considered a non-informative prior distribution $\pi(\phi)$ for ϕ, while they assumed c to be a fixed constant.

2.2 Bayes factors.

The Bayes factor (Jeffreys, 1961) is used in Bayesian hypotheses testing and model comparison. Assume that $f_i(\mathbf{x}|\theta_i)$, $i = 1, 2$ are the likelihoods for $\mathbf{x}$ under two competing models M_1 and M_2, with parameters θ_i, and let $\pi_i(\theta_i)$ be their prior distributions. The Bayes factor for M_2 against M_1

$$B_{21} = \frac{\int f_2(\mathbf{x}|\theta_2)\pi_2(\theta_2)d\theta_2}{\int f_1(\mathbf{x}|\theta_1)\pi_1(\theta_1)d\theta_1} \ , \tag{3}$$

represents the weight of evidence from the data in favor of M_2, against M_1. The Bayes factor (3) is indeterminate when non-informative improper prior distributions are considered, as in the assumptions made by CCO for the models of Section 2.1. Arguments have been made by a number of authors (Aitkin, 1991; O'Hagan, 1995; Berger and Pericchi, 1996a, b; De Santis and Spezzaferri, 1999) suggesting the use of the partial Bayes factor when weak prior information is represented by improper distributions. The partial Bayes factor is defined by:

$$B_{21}^P(\mathbf{x}(\ell)) = \frac{\int f_2(\mathbf{x}|\theta_2)\pi_2(\theta_2)d\theta_2}{\int f_1(\mathbf{x}|\theta_1)\pi_1(\theta_1)d\theta_1} \frac{\int f_1(\mathbf{x}(\ell)|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(\mathbf{x}(\ell)|\theta_2)\pi_2(\theta_2)d\theta_2} \ ,$$

where a subset $\mathbf{x}(\ell)$, $0 < \ell < n$, of the data is used as a training sample for updating the priors into proper posterior distributions and the models are compared using the remainder of data. To avoid arbitrariness in the choice of a specific training sample $\mathbf{x}(\ell)$, O'Hagan (1995) proposed using the whole likelihood raised to the power $b = \ell/n$ for training. This choice was motivated by the following approximation:

$$f_i(\mathbf{x}(\ell)|\theta_i) \approx [f_i(\mathbf{x}|\theta_i)]^b \ , \tag{4}$$

that is valid for iid data, when n and ℓ are large. O'Hagan defined the fractional Bayes factor as:

$$B_{21}^F(\mathbf{x}, b) = \frac{q_2(\mathbf{x}, b)}{q_1(\mathbf{x}, b)}, \tag{5}$$

where

$$q_i(\mathbf{x}, b) = \frac{\int f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i)d\theta_i}{\int [f_i(\mathbf{x}|\theta_i)]^b \pi_i(\theta_i)d\theta_i} \ . \tag{6}$$

A non-asymptotic justification for using (5) was given in De Santis and Spezzaferri (1997). A different method for eliminating the influence of training samples on partial Bayes factor was suggested by Berger and Pericchi (1996a) and led to the intrinsic Bayes factor approach.

For the models in (1), under the prior assumptions of Section 2.1, CCO (2000) obtained the following expressions for q_i in (6):

$$\begin{aligned} q_1(\mathbf{x}, b) &= \frac{\int \prod_{j=0}^{k} \alpha_j(\phi)^{r_j} \pi_1(\phi) d\phi}{\int \prod_{j=0}^{k} \alpha_j(\phi)^{br_j} \pi_1(\phi) d\phi} \\ q_2(\mathbf{x}, b) &= \frac{\Gamma(bn+c)}{\Gamma(n+c)} \frac{\int \prod_{j=0}^{k} \frac{\Gamma(r_j + c\alpha_j(\phi))}{\Gamma(c\alpha_j(\phi))} \pi_2(\phi) d\phi}{\int \prod_{j=0}^{k} \frac{\Gamma(br_j + c\alpha_j(\phi))}{\Gamma(c\alpha_j(\phi))} \pi_2(\phi) d\phi} \ , \end{aligned} \tag{7}$$

and they used the fractional Bayes factor (5) for Bayesian goodness of fit testing. In CCO′s paper the expression of q_2 in (7) and, consequentely, the value of the fractional Bayes factor depend on the values chosen for the Dirichlet parameter c. This left some ambiguities for a clear interpretation of the goodness of fit test proposed by these authors.

2.3 The generalized fractional Bayes factor.

We examine here a proposal from De Santis and Spezzaferri (1997, 2001) that leads to a variation of the alternative Bayes factors used in model comparison when the prior information is weak. Like the fractional and the intrinsic Bayes factors, the method avoids arbitrary choices of specific training samples. This method is a generalization of the fractional Bayes factor and we will show that there are cases of model comparisons in the Bayesian framework where this method outperforms the fractional Bayes factor.

De Santis and Spezzaferri (1997, 2001) suggested replacing the likelihood $f_i(\mathbf{x}(\ell)|\theta_i)$ with the geometric training likelihood:

$$\mathcal{L}_i^{\ell}(\theta_i) = \Big[\prod_{\mathbf{x}(\ell) \in \mathcal{X}} f_i(\mathbf{x}(\ell)|\theta_i) \Big]^{\frac{1}{L}}, \tag{8}$$

where $\mathcal{X} = \{\mathbf{x}_1(\ell), \cdots \mathbf{x}_h(\ell) \cdots \mathbf{x}_L(\ell)\}$ is the set of all possible training samples of size ℓ and $L \leq \binom{n}{\ell}$ is the total number of samples considered. The geometric training likelihood $\mathcal{L}_i^{\ell}(\theta_i)$ reduces to $[f_i(\mathbf{x}(\ell)|\theta_i)]^b$, $(b = \ell/n)$ when observations are iid (De Santis and Spezzaferri, 2001), but when data are either dependent or not identically distributed or both, the geometric training likelihood (8) can be used for defining the *generalized fractional Bayes factor*:

$$B_{21}^{GF} = \frac{q_2^G(\mathbf{x}, \ell)}{q_1^G(\mathbf{x}, \ell)}, \tag{9}$$

where:

$$q_i^G(\mathbf{x}, \ell) = \frac{\int f_i(\mathbf{x}|\theta_i)\ \pi(\theta_i)\ d\theta_i}{\int \mathcal{L}_i^{\ell}(\theta_i)\ \pi(\theta_i)\ d\theta_i} \ .$$

De Santis and Spezzaferri (2001) showed that, when comparing nested linear models, with independent but not identically distributed data, the generalized fractional Bayes factor always retains consistency, while the fractional Bayes factor may not. Lack of consistency in a model choice criterion implies that for increasing sample size the probability of choosing the correct model does not approach to 1.

2.4 Advantages of the Generalized Fractional Bayes factor

This section presents some examples of model comparison where the generalized fractional Bayes factor B_{21}^{GF} behaves more reasonably than the fractional Bayes factor B_{21}^{F}.

Example 1. Consider the two equicorrelation models:

$$\begin{aligned} M_1: \ \mathbf{x}|\rho \ &\sim \mathbb{N}\big(\mathbf{0}, \ (1-\rho)\mathbf{I} + \rho\mathbf{J}\big) \\ M_2: \ \mathbf{x}|\rho,\mu \ &\sim \mathbb{N}\big(\mu\mathbf{1}, (1-\rho)\mathbf{I} + \rho\mathbf{J}\big) \end{aligned}$$

where the value of the correlation ρ is assumed known and the only unknown parameter is the mean μ in M_2. The vector $\mathbf{x}$ consists of n observations, and $\mathbf{0}, \mathbf{1}, \mathbf{I}, \mathbf{J}$ are, respectively, vectors of n zeros, n ones, the $n \times n$ identity matrix and an $n \times n$ matrix of ones.

Let the prior distribution for μ be improper, $\pi(\mu) \propto$ constant, and examine the expressions of fractional and generalized fractional Bayes factors for M_2 against M_1, with $\ell = 1$:

$$\begin{aligned} B_{21}^{F} \ &= \frac{1}{\sqrt{n}} \ \exp\Big\{+\frac{1}{2} \ \frac{(n-1)\bar{x}^2}{1+(n-1)\rho}\Big\} \\ B_{21}^{GF} \ &= \frac{\sqrt{1+(n-1)\rho}}{\sqrt{n}} \ \exp\Big\{+\frac{1}{2} \ \frac{(n-1)(1-\rho)\bar{x}^2}{1+(n-1)\rho}\Big\}. \end{aligned}$$

When $\rho = 0$, B_{12}^{F} and B_{12}^{GF} coincide, since observations are independent and $\mathcal{L}_i^{\ell}(\theta_i) = [f_i(\mathbf{x}(\ell)|\theta_i)]^b$. When, instead $\rho \neq 0$, the data are strongly dependent. No Bayes factor is a consistent model selector since $\bar{x}$ is not even a consistent estimator of μ. However, we see that $B_{21}^{F} \xrightarrow{p} 0$ whether or not M_1 and M_2 is true. In contrast, B_{21}^{GF} stays bounded in probability away from zero and its distribution is stochastically larger under M_2 than M_1.

Thus for large n, whatever the evidence from data, the fractional Bayes factor will always choose the null model M_1. The generalized fractional Bayes factor, instead, is asymptotically equivalent to $\sqrt{\rho} \ \exp\{(2\rho)^{-1}(1-\rho)\bar{x}^2\}$ and it allows one to make a reasonable choice between M_1 and M_2, based on the distance of $\bar{x}$ from 0. It can be shown that the behavior of fractional Bayes factor is due to the fact that, under M_2, the fraction of the likelihood that is supposed to get information about μ fails to depend on μ when n is large enough.

Example 2. This example shows another case of model comparison with dependent data, where the generalized fractional Bayes factor appears to perform more reasonably than the fractional Bayes factor. Consider the following autoregressive models of order 1:

$$\begin{aligned} M_1: \ \mathbf{x}|\rho \ &\sim \mathbb{N}\Big(\mathbf{0}, \ \frac{1}{1-\rho^2}\mathbf{W}\Big) \\ M_2: \ \mathbf{x}|\rho,\mu \ &\sim \mathbb{N}\Big(\mu\mathbf{1}, \frac{1}{1-\rho^2}\mathbf{W}\Big) \ , \end{aligned}$$

where the matrix $\mathbf{W} = \{w_{ss'}; s, s' = 1, \cdots n\}$ has elements $w_{ss'} = \rho^{|s-s'|}$ and ρ is a known constant. Let the prior distribution for μ be $\pi(\mu) \propto$ constant. Setting $\ell = 1$, the two fractional Bayes factors for M_2 against M_1 are:

$$\begin{aligned} B_{21}^{F} &= \frac{1}{\sqrt{n}} \ \exp\Big\{+\frac{(1-\rho)}{2} \ \frac{(n-1)(n\bar{x}-\rho S)^2}{n[n-\rho(n-2)]}\Big\} \\ B_{21}^{GF} &= \frac{\sqrt{1+\rho}}{\sqrt{n-\rho(n-2)}} \ \exp\Big\{+\frac{(1-\rho)}{2} \ \Big[\frac{(n\bar{x}-\rho S)^2}{n-\rho(n-2)} - (1+\rho)\bar{x}^2\Big]\Big\}, \end{aligned} \tag{10}$$

where $S = \sum_{s=2}^{n-1} x_s$. Both Bayes factors in (10) are consistent and, as in the previous example, they coincide when $\rho = 0$. Consider now the limiting case when, keeping everything else fixed, ρ increases up to 1, that is equivalent to dealing with a single observation, no matter what the value of n is. In this case:

$$\lim_{\rho\to 1} B_{21}^{F} = \frac{1}{\sqrt{n}} \qquad\qquad \lim_{\rho\to 1} B_{21}^{GF} = 1 \ .$$

Thus, for increasing ρ, when data become more and more dependent, the fractional Bayes factor will favor M_1 whatever the data. The generalized fractional Bayes factor instead suggests, more appropriately, that no choice can be made among the two models in this limiting case.

Next we compare the two fractional Bayes factors when hierarchical priors are used. Let data be iid and let the prior distribution on the parameter of one of the two models, θ_2 in M_2 say, be set in two stages: $\pi_2(\theta_2|\psi_2)$ and $\pi_2(\psi_2)$, where ψ_2 is a hyperparameter on which prior information is weak. In this case the expressions of the fractional and of the generalized fractional Bayes factors coincide:

$$B_{21}^{F} = B_{21}^{GF} = \frac{\int\int f_2(\mathbf{x}|\theta_2)\pi_2(\theta_2|\psi_2)\pi_2(\psi_2)d\psi_2 d\theta_2}{\int\int [f_2(\mathbf{x}|\theta_2)]^b \pi_2(\theta_2|\psi_2)\pi_2(\psi_2)d\psi_2 d\theta_2} \ \frac{\int [f_1(\mathbf{x}|\theta_1)]^b \pi_1(\theta_1)d\theta_1}{\int f_1(\mathbf{x}|\theta_1)\pi_1(\theta_1)d\theta_1} \ . \tag{11}$$

But, it is known that when a hierarchical two-stage prior is considered, the prior at the first stage can be collapsed with the likelihood to produce the marginal distribution $f_2(\mathbf{x}|\psi_2)$ of data given ψ_2. Then, under M_2, one can take $f_2(\mathbf{x}|\psi_2) = \int f_2(\mathbf{x}|\theta_2)\pi_2(\theta_2|\psi_2)d\theta_2$ to be the data distribution, conditional on the parameter ψ_2 with prior distribution $\pi_2(\psi_2)$. Thus, differently from the procedures presented so far, when a hierarchical prior distribution is chosen, we propose the relevant Bayes factors for model selection be built using $f_2(\mathbf{x}|\psi_2)$ as data distribution under M_2 and $\pi_2(\psi_2)$ as prior distribution. Typically, the observations $\mathbf{x}$, with density $f_2(\mathbf{x}|\psi_2)$, are dependent, in which case the fractional Bayes factor in (5) cannot be computed with $f_2(\mathbf{x}|\phi_2)$, since the approximation (4) is valid only for iid data. Thus, when hierarchical priors are considered, the appropriate fractional Bayes factor is the one defined in (11). In contrast, the generalized fractional Bayes factor (9) takes the following form

$$\tilde{B}_{21}^{GF} = \frac{\int f_2(\mathbf{x}|\psi_2) \ \pi_2(\psi_2) \ d\psi_2}{\int \mathcal{L}_2^{\ell}(\psi_2) \ \pi_2(\psi_2) \ d\psi_2} \ \frac{\int [f_1(\mathbf{x}|\theta_1)]^b \ \pi_1(\theta_1) \ d\theta_1}{\int f_1(\mathbf{x}|\theta_1) \ \pi_1(\theta_1) \ d\theta_1}, \tag{12}$$

where $\mathcal{L}_2^\ell(\psi_2) = \left[\prod_{\mathbf{x}(\ell)} f_2(\mathbf{x}(\ell)|\psi_2)\right]^{\frac{1}{L}}$, and $f_2(\mathbf{x}(\ell)|\psi_2) = \int f_2(\mathbf{x}(\ell)|\theta_2)\ \pi_2(\theta_2|\psi_2)\ d\theta_2$.

We now illustrate an argument in favor of using $\tilde{B}_{21}^{GF}$ instead of B_{21}^F for comparing two models M_1 and M_2, when the prior distribution is set through a two-stage hierarchy for M_2. The following Lemma 1 shows that the fractional Bayes factor B_{21}^F gives more support to M_2 than the generalized fractional Bayes factor $\tilde{B}_{21}^{GF}$, regardless of which models is true. Lemma 2 extends this result and shows that the fractional Bayes factor supports M_2 more strongly when information from data on θ_2 are in disagreement. This supports our point that the generalized fractional Bayes factor is to be preferred to the the fractional Bayes factor.

Lemma 1 *In the hierarchical prior setup specified above, the fractional Bayes factor B_{21}^F gives more support to M_2 than the generalized fractional Bayes factor $\tilde{B}_{21}^{GF}$ does:*

$$\frac{\tilde{B}_{21}^{GF}}{B_{21}^F} = \frac{\int\int \left[f_2(\mathbf{x}|\theta_2)\right]^{\ell/n} \pi_2(\theta_2|\psi_2)\pi(\psi_2)\ d\theta_2\ d\psi_2}{\int \mathcal{L}_2^\ell(\psi_2)\pi(\psi_2)\ d\psi_2} \leq 1\ . \tag{13}$$

Proof: Use the fact that data, distributed as $f_2(\mathbf{x}|\theta_2)$, are independent and the Cauchy-Schwartz inequality

$$\int \Big[f_2(\mathbf{x}|\theta_2)\Big]^{\frac{\ell}{n}} \pi_2(\theta_2|\psi_2)\ d\theta_2 = \int \Big[\prod_{\mathbf{x}(\ell)\in\mathcal{X}} f_2(\mathbf{x}(\ell)|\theta_2)\Big]^{\frac{1}{L}} \pi_2(\theta_2|\psi_2)\ d\theta_2 \leq \mathcal{L}_2^\ell(\psi_2),$$

and (13) follows. ■

Consider now the definition of *contradictory evidence* on a parameter η given by two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ from a random variable X with density function $f(x|\eta)$.

Definition 1 *We say that two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ give "contradictory evidence" about η if:*

$$\frac{\int f(\mathbf{x}_1|\eta)\ \pi(\eta|\mathbf{x}_2)\ d\eta}{\int f(\mathbf{x}_1|\eta)\ \pi(\eta)\ d\eta} \leq 1. \tag{14}$$

Inequality (14) obtains when $\pi(\eta|\mathbf{x}_2)$, the posterior distribution of η given $\mathbf{x}_2$, puts more weight than the prior distribution $\pi(\eta)$ on values of η where the likelihood $f(\mathbf{x}_1|\eta)$ is small. This shows that the two samples contain opposite (or *contradictory*) information about η.

We now extend Definition 1 to the set $\mathcal{X} = \{\mathbf{x}_1(\ell), \cdots, \mathbf{x}_h(\ell), \cdots, \mathbf{x}_L(\ell)\}$ of all training samples of size ℓ extracted from a sample $\mathbf{x}$ with density $f(\mathbf{x}|\eta)$.

Definition 2 *Let $\pi(\eta|\mathbf{x}_1(\ell), \cdots \mathbf{x}_{h-1}(\ell))$ denote the posterior distribution for η obtained from the first $(h-1)$ training samples. We say that the information about η from the h-th training sample is "contradictory" with respect to the previous $(h-1)$ training samples if:*

$$\frac{\int f(\mathbf{x}_h(\ell)|\eta)\ \pi(\eta|\mathbf{x}_1(\ell), \ldots \mathbf{x}_{h-1}(\ell))\ d\eta}{\int f(\mathbf{x}_h(\ell)|\eta)\ \pi(\eta)\ d\eta} \leq 1. \tag{15}$$

Lemma 2 *The larger is the number of training samples with contradictory evidence about θ_2 as defined in (15), the smaller is the ratio $\tilde{B}_{21}^{GF}/B_{21}^{F}$ in (13)*

Proof: From Jensen's inequality:

$$\int \Big[f_2(\mathbf{x}|\theta_2)\Big]^{\frac{\ell}{n}} \pi_2(\theta_2|\psi_2)\ d\theta_2 \leq \left[\int \prod_{\mathbf{x}(\ell)\in\mathcal{X}} f_2(\mathbf{x}(\ell)|\theta_2)\ \pi_2(\theta_2|\psi_2)\ d\theta_2\right]^{\frac{1}{L}}, \tag{16}$$

and, using the Bayes theorem recursively, the integral on the right hand side of (16) can be written as:

$$\int \prod_{\mathbf{x}(\ell)\in\mathcal{X}} f_2(\mathbf{x}(\ell)|\theta_2)\ \pi_2(\theta_2|\psi_2)\ d\theta_2 = \prod_{h=1}^{L} \int f_2(\mathbf{x}_h(\ell)|\theta_2)\ \pi_2\big[\theta_2|\mathbf{x}_1(\ell),\cdots\mathbf{x}_{h-1}(\ell),\psi_2\big]\ d\theta_2. \tag{17}$$

Using (16) and (17) we obtain:

$$\frac{\int \big[f_2(\mathbf{x}|\theta_2)\big]^{\ell/n}\pi_2(\theta_2|\psi_2)\ d\theta_2}{\mathcal{L}_2^{\ell}(\psi_2)} \leq \left[\prod_{h=1}^{L} \frac{\int f_2(\mathbf{x}_h(\ell)|\theta_2)\ \pi_2(\theta_2|\mathbf{x}_1(\ell),\cdots\mathbf{x}_{h-1}(\ell),\psi_2)\ d\theta_2}{\int f_2(\mathbf{x}_h(\ell)|\theta_2)\ \pi_2(\theta_2|\psi_2)\ d\theta_2}\right]^{\frac{1}{L}}. \tag{18}$$

The h-th term ($h = 1,\cdots L$) in the product on the right hand side of (18) is less than 1 when information about plausible values of θ_2 among the first $h-1$ training samples and the h-th is contradictory as defined in (15). Thus the product on the right hand side of (18) turns out to be much smaller than 1 whenever there is a large number of training samples with contradictory information on θ_2. ■

Lemma 2 shows that the generalized fractional Bayes factor $\tilde{B}_{21}^{GF}$ has a smaller chance than B_{21}^{F} of choosing a model on whose parameter the sample evidence is contradictory. We argue that this is the appropriate behavior for a model choice criterion.

3. Hierarchical prior specified through a Dirichlet process.

3.1 Introduction.

Let us now go back to the model selection problem of Section 2.1. We assume, as in CCO, that the prior distribution of $\boldsymbol{\alpha}$ under M_2 is set in two stages. At the first stage, given ϕ and c, let $\boldsymbol{\alpha}$ be distributed as in (2), with $c_j = c\alpha_j(\phi)$, $j = 0,\cdots k$. At the second stage, in contrast to CCO, we assume that $\pi_2(\phi, c)$ is a noninformative prior distribution for both hyperparameters ϕ and c. Deriving the marginal distribution $m_2(\mathbf{x}|\phi, c)$ for M_2, the two models to be compared are:

$$\begin{aligned} M_1: \quad m_1(\mathbf{x}|\phi) &= \prod_{j=0}^{k} [\alpha_j(\phi)]^{r_j} \\ M_2: \quad m_2(\mathbf{x}|\phi, c) &= \frac{\prod_{j=0}^{k}\prod_{h=0}^{r_j-1}(c\alpha_j(\phi)+h)}{\prod_{h=0}^{n-1}(c+h)}\ . \end{aligned} \tag{19}$$

In (19), we take the second product in the numerator of $m_2(\mathbf{x}|\phi,c)$ to be 1 if $r_j = 0$. Data under M_2 are not iid, thus we consider the generalized fractional Bayes factor approach of Section 2.3 for model selection. The two models in (19) are nested, in fact $\lim_{c\to\infty} m_2(\mathbf{x}|\phi,c) = m_1(\mathbf{x}|\phi)$. We interpret $c > 0$ as a parameter measuring the distance between the two models, since the smaller c is, the further apart M_2 is from M_1.

The generalized fractional Bayes factor for the models in (19) is:

$$\tilde{B}_{21}^{GF} = \frac{\int m_2(\mathbf{x}|\phi,c)\pi_2(\phi,c)d\phi dc}{\int \mathcal{L}_2^{\ell}(\phi,c)\pi_2(\phi,c)d\phi dc} \, \frac{\int [m_1(\mathbf{x}|\phi)]^{\frac{\ell}{n}} \pi_1(\phi)d\phi}{\int m_1(\mathbf{x}|\phi)\pi_1(\phi)d\phi} \; . \tag{20}$$

Lemma 3 *The explicit expression for $\mathcal{L}_2^{\ell}(\phi,c)$ in (20) is given by:*

$$\mathcal{L}_2^{\ell}(\phi,c) = \frac{\prod_{j=0}^{k}\prod_{h=0}^{\min(r_j,\ell)-1}(c\alpha_j(\phi)+h)^{\beta_{h,j}}}{\prod_{h=0}^{\ell-1}(c+h)} \; , \tag{21}$$

where

$$\beta_{h,j} = \frac{1}{L}\sum_{t=\max(h+1,\ell+r_j-n)}^{\min(r_j,\ell)} \binom{r_j}{t}\binom{n-r_j}{\ell-t}.$$

The sum on the right hand side of $\beta_{h,j}$ is the total number of training samples with at least $h+1$ observations in group G_j.

Proof: From (19) the likelihood of one training sample $\mathbf{x}(\ell) \in \mathcal{X}$ under M_2 is:

$$m_2(\mathbf{x}(\ell)|\phi,c) = \frac{\prod_{j=0}^{k}\prod_{h=0}^{r_j(\mathbf{x}(\ell))-1}(c\alpha_j(\phi)+h)}{\prod_{h=0}^{\ell-1}(c+h)} \; , \tag{22}$$

where $r_j(\mathbf{x}(\ell))$ is the number of observations in $\mathbf{x}(\ell)$ falling in group G_j and, as in (19), if $r_j(\mathbf{x}(\ell)) = 0$ the second product in the numerator is set to 1. From (8) and (22), $\mathcal{L}_2^{\ell}(\phi,c)$ is given by:

$$\left[\prod_{\mathbf{x}(\ell)\in\mathcal{X}} \frac{\prod_{j=0}^{k}\prod_{h=0}^{r_j(\mathbf{x}(\ell))-1}(c\alpha_j(\phi)+h)}{\prod_{h=0}^{\ell-1}(c+h)}\right]^{\frac{1}{L}} = \frac{\left[\prod_{j=0}^{k}\prod_{\mathbf{x}(\ell)\in\mathcal{X}}\prod_{h=0}^{r_j(\mathbf{x}(\ell))-1}(c\alpha_j(\phi)+h)\right]^{\frac{1}{L}}}{\prod_{h=0}^{\ell-1}(c+h)} \; . \tag{23}$$

For all $\mathbf{x}(\ell) \in \mathcal{X}$, we have $r_j(\mathbf{x}(\ell)) \leq \min(r_j,\ell)$. Now let $I[r_j(\mathbf{x}(\ell)) - 1 \geq h]$ denote the indicator function of the event $\{r_j(\mathbf{x}(\ell)) - 1 \geq h\}$, so the third product in the numerator of (23) can be written as:

$$\prod_{h=0}^{r_j(\mathbf{x}(\ell))-1}(c\alpha_j(\phi)+h) = \prod_{h=0}^{\min(r_j,\ell)-1}(c\alpha_j(\phi)+h)^{I[r_j(\mathbf{x}(\ell))-1\geq h]} \tag{24}$$

Substituting (24) in (23) and appropriately adding the indicator functions in (24), we obtain:

$$\mathcal{L}_2^\ell(\phi, c) = \frac{\left[\prod_{j=0}^{k} \prod_{h=0}^{\min(r_j,\ell)-1} (c\alpha_j(\phi) + h)^{\sum_{\mathbf{x}(\ell)} I[r_j(\mathbf{x}(\ell))-1 \geq h]}\right]^{\frac{1}{L}}}{\prod_{h=0}^{\ell-1}(c+h)}$$

from which (21) follows. ■

From Lemma 3 it can be checked that when $\ell = 1$, $\mathcal{L}_2^\ell(\phi, c)$ reduces to $[m_1(\mathbf{x}, \phi)]^{\ell/n}$ thus implying that if the size of training samples is $\ell = 1$ the geometric likelihood $\mathcal{L}_2^\ell(\phi, c)$ fails to update the prior on c, as it was to be expected since, in the case considered here, the size of the minimal training sample is 2.

Remark So far, the number of groups G_j has been assumed to be finite. In the case where observations fall in one of a countable set of groups, as for example in the Poisson case, then in M_2 we assume that $\mathbb{P}(x_s \in G_j|\phi) = \alpha_j = F(j) - F(j-1)$, where $F(\cdot)$ is a discrete time Dirichlet process, centered on M_1, with parameter $cH(\cdot|\phi)$, where $H(\cdot|\phi)$ is a cumulative distribution function with jumps $\alpha_j(\phi)$ at the points $j = 0, 1, \cdots$. If $k-1$ denotes the last group index where observations fall, then results and notation of this Section still hold with $\alpha_k(\phi) = 1 - \sum_{j=0}^{k-1} \alpha_j(\phi)$, and $r_k = 0$.

3.2 Prior distributions for ϕ and c.

CCO treated c as a known constant and chose the same noninformative prior distribution $\pi_1(\phi)$ for ϕ under both models. The form of $\pi_1(\phi)$ was obtained according with the specific parametric family $\mathcal{F} = \{\alpha_j(\phi),\ \phi \in \mathbb{R}\}$ considered.

We will now specify a prior $\pi_1(\phi)$ in M_1 and a joint prior $\pi_2(\phi, c)$ in M_2. An independence assumption among ϕ and c under M_2 appears reasonable, given that c represents the distance among M_1 and M_2 in (19), and ϕ specifies the location of the two models in (19). Hence $\pi_2(\phi, c) = \pi_2(\phi)\ \pi_2(c)$, and we choose $\pi_2(\phi) = \pi_1(\phi)$, as in CCO. The standard noninformative prior distribution for the scale parameter c, $\pi_2(c) \propto 1/c$ cannot be used in this context, since the integrals in (20) do not converge. This problem has been considered in the literature by Good (1967), who suggested two possible solutions. Instead of looking for a proper distribution approximating $\pi_2(c) \propto 1/c$ as in Good, we propose an approach that exploits the fact that the correlation among any two observations under M_2 is a function of c. From $m_2(\mathbf{x}|\phi, c)$ in (19):

$$r(x_s, x_{s'}) = \frac{\mathbb{E}[x_s x_{s'}] - \mathbb{E}[x_s]\ \mathbb{E}[x_{s'}]}{\sqrt{\mathbb{V}[x_s]\ \mathbb{V}[x_{s'}]}} = \frac{1}{c+1},$$

showing that the correlation among two observations does not depend on ϕ, and it varies between 0 and 1. If no prior information is available for $r(x_s, x_{s'})$, we assume $r(x_s, x_{s'})$ to be uniformly distributed between 0 and 1, and obtain from this the correspondent proper prior distribution for c

$$\pi_2(c) \propto \frac{1}{(c+1)^2}. \tag{25}$$

3.3 Examples and comparisons with CCO′s procedure.

The fractional Bayes factor in (5) and the generalized fractional Bayes factor in (20) can be computed once the parametric family $\mathcal{F} = \{\alpha_j(\phi),\ \phi \in \mathbb{R}\}$ that specifies $\alpha_j(\phi)$ under M_1 is selected. We will consider the case where the $\alpha_j(\phi)'$s have the form of a binomial distribution, so that

$$\alpha_j(\phi) = \binom{k}{j}\phi^j(1-\phi)^{k-j} \qquad j = 0, 1, \ldots, k \tag{26}$$

and, as in CCO, we choose the improper prior distribution for ϕ to be $\pi_1(\phi) \propto \phi^{-1}(1-\phi)^{-1}$. We first examine data from Hoaglin *et al.* (1985), giving the number of females in 100 queues of size 10 observed in a London underground station, also examined by CCO to test the hypothesis that data were in fact from a binomial distribution. The data are below.

Number of females	0	1	2	3	4	5	6	7	8	9	10
Number of queues	1	3	4	23	25	19	18	5	1	1	0

The fractional Bayes factor computed by CCO depends on the parameter c. The values obtained are not stable and they get closer to 1 for large c when in fact the two models M_1 and M_2 coincide. Table 1 shows the values of the fractional Bayes factor reported by CCO for a few values of c and with $\ell = 1$, together with the value of the generalized fractional Bayes factor computed from (21), and based on training samples of size $\ell = 2$.

B_{21}^F			
$c = 2$	$c = 20$	$c = 100$	$\tilde{B}_{21}^{GF}$
8.0×10^{-7}	0.355	0.676	4.68×10^{-6}

TABLE 1

The value obtained for the generalized Bayes factor in this example supports the conclusion from the fractional Bayes factor, but avoids the need to guess which value of c might be appropriate and why.

In the example above, the sample size was $n = 100$. For further comparison, we selected a few examples with decreasing sample size, and data generated either from a binomial distribution, or from an U shaped distribution. Reducing the sample size makes it harder for any model selection method to choose the correct model.

Table 2 shows three data sets generated with $\alpha_j(\phi)$ defined in (26) with $k = 6$, $\phi = .5$ and decreasing sample size:

								B_{21}^F			
n	r_0	r_1	r_2	r_3	r_4	r_5	r_6	$c = 2$	$c = 20$	$c = 100$	$\tilde{B}_{21}^{GF}$
60	0	3	14	22	15	6	0	0.0775	0.7437	1.0954	3.94×10^{-5}
40	0	4	7	19	8	1	1	0.1225	1.6461	1.6419	0.0002
20	1	1	8	5	3	2	0	0.0493	0.7192	1.0655	0.0002

TABLE 2

Similarly Table 3 considers three data sets generated with $\alpha_j(\phi)$ defined in (26) with $k = 6$, $\phi = .25$ and decreasing sample size:

								B_{21}^F			
n	r_0	r_1	r_2	r_3	r_4	r_5	r_6	$c = 2$	$c = 20$	$c = 100$	$\tilde{B}_{21}^{GF}$
60	10	22	22	6	0	0	0	0.4755	1.5874	1.4471	4.35×10^{-5}
40	11	10	12	4	3	0	0	0.1540	1.3050	1.5214	6.54×10^{-5}
20	2	10	3	3	1	1	0	0.1520	1.1507	1.3288	0.0004

TABLE 3

The main advantage in using the generalized fractional Bayes factor is that it avoids the instability caused by the arbitrary choice of the Dirichlet parameter c. The results obtained in Lemmas 1 and 2, also appear to affect B_{21}^F in these examples, since model M_2 is more likely to be chosen. This is even clearer in Table 3, where data generated from an asymmetric binomial with small n, are also likely to be from a non-binomial model.

For checking how well the fractional Bayes factors detect data that are sampled from a distribution that is obviously not binomial, Table 4 considers three data sets with decreasing values of n, $k = 5$ and data sampled from $\boldsymbol{\alpha} = (0.31, 0.16, 0.03, 0.03, 0.16, 0.31)$ representing a U shaped distribution. Table 4 shows that the fractional Bayes factor is able to detect better than $\tilde{B}_{21}^{GF}$ that data are not binomial, in accordance with our general results of Section 2.

							B_{21}^F			
n	r_0	r_1	r_2	r_3	r_4	r_5	$c = 2$	$c = 20$	$c = 100$	$\tilde{B}_{21}^{GF}$
60	15	10	3	4	8	20	7.73×10^{24}	9.92×10^{21}	1.56×10^{14}	1.73×10^{19}
40	15	6	0	0	5	14	2.68×10^{26}	2.32×10^{21}	9.58×10^{13}	2.49×10^{20}
20	7	3	0	0	3	7	2.69×10^{11}	9.9×10^{7}	$10,977.76$	$207,570.6$

TABLE 4

The case where the $\alpha_j(\phi)$'s have the form of a Poisson distribution leads to similar results that are not reported here.

4. Extension to a Beta-Stacy process.

This section generalizes the approach of Section 3. We still consider the two models $M_1 : \mathbb{P}(x_s \in G_j|\phi) = \alpha_j(\phi)$ and $M_2 : \mathbb{P}(x_s \in G_j|\boldsymbol{\alpha}) = \alpha_j$, where under M_1 the α_js belong to a specific parametric family $\mathcal{F} = \{\alpha_j(\phi),\ \phi \in \mathbb{R}\}$ with the number of groups G_j being either finite or countable. In particular, we explain how the method can be extended to the Beta-Stacy process, which is a generalization of the Dirichlet.

As in Section 3, the prior distribution for $\boldsymbol{\alpha}$ under M_2 is set in two stages. We assume that $\boldsymbol{\alpha}$ is a realization of a discrete Beta-Stacy process, with appropriate hyperparameters on which a second stage prior distribution is specified.

A Beta-Stacy process is a generalization of the Dirichlet process. The main features of the Beta-Stacy distribution and discrete process, as obtained by Walker and Muliere (1997) are summarized in the Appendix.

The first stage prior distribution for α_j in M_2 considers α_j to be the width of the j-th jump $F(j)-F(j-1)$ of a discrete Beta-Stacy process $F(\cdot)$ as in Definition A1. We constrain the discrete Beta-Stacy process parameters $\{\delta_j, \beta_j\}$ to be such that the model M_2 is centered on M_1 thus, as in Section 3, $\mathbb{E}[\alpha_j] = \alpha_j(\phi)$.

Let $H(\cdot|\phi)$ be the cumulative distribution function specified by the parametric family $\mathcal{F}$ in M_1. From Lemma A1, M_2 is centered on M_1 if $\delta_j = c_j(H(j|\phi) - H(j-1|\phi))$ and $\beta_j = c_j(1 - H(j|\phi))$, where $c_j > 0$. By requiring M_2 be centered on M_1 reduces the number of hyperparameters of the first stage prior to ϕ and $\mathbf{c} = (c_0, c_1, \cdots, c_j, \cdots)$.

As in Section 3 we can compare M_1 and M_2 using the generalized fractional Bayes Factor $\tilde{B}_{21}^{GF}$ defined in (12). The Beta-Stacy assumptions make it cumbersome to write the expression for $m_2(\mathbf{x}|\phi, \mathbf{c})$ explicitely as in (19) where a Dirichlet prior was considered. Using the relation among the conditional densities of the elements in $\mathbf{x}$, the following expression for $m_2(\mathbf{x}|\phi, \mathbf{c})$ obtains

$$m_2(\mathbf{x}|\phi, \mathbf{c}) = m_2(x_1|\phi, \mathbf{c}) m_2(x_2|x_1, \phi, \mathbf{c}) \cdots m_2(x_n|x_1, \cdots x_{n-1}, \phi, \mathbf{c}), \tag{27}$$

where each factor in (27) can be obtained through (33), the predictive probability derived in Lemma A3.

The marginal distribution of each training sample $\mathbf{x}(\ell) \in \mathcal{X}$ needed to calculate the generalized fractional Bayes factor can be obtained similarly to (27).

Note that, denoting by G_k the last group containing observations we have that $m_2(\mathbf{x}|\phi, \mathbf{c})$ depends only on the hyperparameters ϕ and $c_0, c_1, \ldots, c_k$, on which we need to specify a prior distribution.

A further reduction of the number of hyperparameters is obtained requiring that the distribution for $\boldsymbol{\alpha}$ is such that the correlation $\rho(\alpha_s, \alpha_r)$ among any two parameters be negative. This condition is an appropriate default requirement because of the constraints $\sum \alpha_j = 1$ and it is satisfied by the Dirichlet prior of Section 3.

Lemma 4 *A sufficient condition for $\rho(\alpha_s, \alpha_r)$ to be negative for all $s \neq r$ is:*

$$c_0 \leq c_1 \leq \cdots \leq c_j \leq \ldots \leq c_0 + 1. \tag{28}$$

Proof:
Lemma A2 in the appendix can be used for deriving the expressions of the covariances of the random probability masses $\{\alpha_0, \alpha_1, \ldots, \alpha_k\}$ assigned to the groups $\{G_0, G_1, \ldots, G_k\}$ by a Beta-Stacy process centered on $H(\cdot|\phi)$. Assuming $s < t$, we have:

$$\alpha_t = V_t \prod_{j=0}^{t-1}(1 - V_j) = \alpha_s \frac{V_t}{V_s} \prod_{j=s}^{t-1}(1 - V_j), \qquad \text{and} \qquad \alpha_s \alpha_t = V_s V_t \prod_{j=0}^{s-1}(1 - V_j)^2 \prod_{j=s}^{t-1}(1 - V_j).$$

Thus $\mathbb{E}(\alpha_t \alpha_s)$ can be obtained as product of first and second moments of Beta random variables and the covariance between α_s and α_t is given by:

$$\mathbb{C}(\alpha_s \alpha_t) = \alpha_s(\phi)\alpha_t(\phi) \left[\frac{c_s}{c_0 + 1} \prod_{j=0}^{s-1} \frac{c_j[1 - H(j|\phi)] + 1}{c_{j+1}[1 - H(j|\phi)] + 1} - 1 \right] \tag{29}$$

The proof follows noting that if (28) holds, all the factors in the product of the expression of $\mathbb{C}(\alpha_s, \alpha_t)$ in (29) are < 1. ■

We suggest using the following specification of $c_0, \ldots, c_k$ that satisfies (28) in Lemma 4:

$$c_j = c + \frac{j}{k}\, d, \quad 0 \leq d \leq 1, \quad j = 0, \ldots k. \tag{30}$$

With (30) the prior distribution for $\boldsymbol{\alpha}$ only depends on the hyperparameters ϕ, c, d, where ϕ and c have the same meaning as in Section 3, and d can be interpreted as the distance between the Dirichlet and Beta-Stacy processes. In fact the two processes coincide when $d = 0$.

The prior distribution for $\boldsymbol{\alpha}$ depends on the way groups have been ordered. This makes the Beta-Stacy hierarchical prior best suited for situations where there is a natural ordering for the groups.

The prior distributions for the hyperparameters ϕ, c and d, can be assigned by adopting the same priors of Section 3 for ϕ and c and a uniform $U(0, 1)$ for d. Then the generalized fractional Bayes factor can be computed as in Section 3.

5. Discussion.

We have presented an approach to Bayesian goodness of fit testing for multinomial data. Our method is based on the generalized fractional Bayes factor. The advantages of this method are that it does not rely on asymptotic justifications, it does not require a subjective prior and it behaves well in the examples. We have also shown that the generalized fractional Bayes factor has some advantages over the fractional Bayes factor.

We used a hierarchical prior distributions built either through a Dirichlet or through a Beta-Stacy model, both expressing non-parametric alternatives to a null sub model of the multinomial. We derived a closed form expression for the generalized fractional Bayes factor under the Dirichlet model prior. The extension to the Beta-Stacy case in Section 4 allows for more flexibility in model specifications, but has greater computational complexity.

The Beta-Stacy hierarchical prior is useful in the case that the groups are ordered. However, the Beta-Stacy prior can also be used in the case where there is no such order. This requires symmetrizing the prior by averaging over group orderings. We leave the details for future work.

Acknowledgments.

The authors would like to thank Larry Wasserman for his help in improving a previous version of this paper.

References

Aitkin, M. (1991). Posterior Bayes Factors. *J. R. Statist. Soc.* B **53**, 111-42.

Albert, J. H. (1997). Bayesian Testing and estimation of association in a two-way contingency table. *J. Am. Statist. Assoc.* **92**, 685-93.

Berger, J.O. and Guglielmi, A. (2001). Bayesian testing of a parametric model versus nonparametric alternatives. *J. Am. Statist. Assoc.* **96**, 542-54.

Berger, J.O. and Pericchi, L. (1996a). The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Assoc.* **91**, 109-22.

Berger, J.O. and Pericchi, L. (1996b). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, J.M. Bernardo *et al.* (Eds.), 23-42, Oxford University Press, Oxford.

Conigliani, C., Castro, J.I., and O'Hagan, A. (2000). Bayesian assessment of goodness of fit against nonparametric alternatives. *Canadian J. Statist.* **28**, 327-42.

De Santis, F. and Spezzaferri, F. (1997). Alternative Bayes factors for model selection. *Canadian J. Statist.* **25**, 503-15.

De Santis, F. and Spezzaferri, F. (1999). Methods for robust and default Bayesian model comparison: the fractional Bayes factor approach. *Int. Statist. Rev.* **67**, 267-86.

De Santis, F. and Spezzaferri, F. (2001). Consistent fractional Bayes factor for nested normal linear models. *J. Statist.Plann. Inference* **97**, 305-21.

Good, I.J. (1967). A Bayesian significance test for multinomial distributions. *J. R. Statist. Soc.* B **29**, 399-431.

Hoaglin, D.C., Mosteller, F., and Tukey, J.O. (1985). *Exploring data, tables, trends and shapes*. John Wiley, New York.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, New York.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Am. Statist. Assoc.* **93**, 1451-60.

O'Hagan, A. (1991). Discussion of Aitkin's "Posterior Bayes Factors". *J. R. Statist. Soc.* B **53**, 111-42.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B **57**, 99-138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test* **6**, 101-18.

Robert, C.P. and Rousseau, J. (2003). A Mixture Approach to Bayesian Goodness of Fit. Preprint. Ceremade - Université Paris-Dauphine.

Verdinelli, I. and Wasserman, L. (1998). Bayesian Goodness of fit Testing using Infinite Dimensional Exponential Families. *The Annals of Statistics*. **26**, 1215-1241.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya-urn scheme. *Ann. Statist.*, **25**, 1762-80.

Appendix: The discrete Beta-Stacy process.

All the results in this appendix are taken from Walker and Muliere (1997). A random variable Y is said to have a Beta-Stacy distribution $BetaSt(\delta, \beta, \xi)$, with parameters $\delta > 0$, $\beta > 0$ and $0 < \xi \leq 1$ if Y has the following density function:

$$f(y|\delta, \beta, \xi) = \frac{1}{B(\delta, \beta)} \frac{1}{\xi^{(\delta+\beta-1)}} y^{\delta-1} (\xi - y)^{\beta-1} I_{(0,\xi)}(y), \tag{31}$$

where $B(\delta, \beta)$ is the beta function. For $\xi = 1$ the density in (31) reduces to the density function of a $Beta(\delta, \beta)$ random variable. For $0 < \xi < 1$ the density in (31) is a *Beta* density that assigns mass to the interval $(0, \xi]$, rather than $(0, 1]$. It can be seen that $\mathbb{E}[Y] = \delta\,\xi/(\delta + \beta)$.

Consider the sequence of points $\{0, 1, \cdots j \cdots\}$ and associate a Beta-Stacy random variable with each j:

$$\begin{array}{rcl} Y_0 & \sim & \text{BetaSt } (\delta_0, \beta_0, 1), \\ Y_1|Y_0 & \sim & \text{BetaSt } (\delta_1, \beta_1, 1 - Y_0), \\ \vdots & & \\ Y_j|Y_0, Y_1, \cdots, Y_{j-1} & \sim & \text{BetaSt } (\delta_j, \beta_j, 1 - \sum_{s=0}^{j-1} Y_s), \\ \vdots & & \end{array} \tag{32}$$

where $\delta_j > 0$, $\beta_j > 0$. Let $F(t) = \sum_{j \leq t} Y_j$, Walker and Muliere (1997) showed that, if $\prod_{j=0}^{\infty} [\beta_j/(\delta_j + \beta_j)] = 0$, then, with probability 1, F is a random cumulative distribution function.

Definition A1 (Walker and Muliere, 1997). The random cumulative distribution function F is a discrete Beta-Stacy random process with parameters $\{\delta_j, \beta_j\}$ and jumps of random size $Y_0, Y_1, \ldots Y_j \ldots$ at the points $\{0, 1, \cdots j \cdots\}$.

Walker and Muliere (1997) also proved a number of properties of the discrete Beta-Stacy random process F, among which we list the following three, as lemmas that we use in Section 4.

Lemma A1 *Let $H(\cdot)$ be a discrete cumulative distribution function with jumps at the points $\{0, 1, \cdots j \cdots\}$. If $\delta_j = c_j(H(j) - H(j-1))$ and $\beta_j = c_j(1 - H(j))$, with $c_j > 0$, the discrete Beta-Stacy process F is centered on H, in the sense that $\mathbb{E}[Y_j] = H(j) - H(j-1)$.*

Remark A1 If $c_j = c$ for all js, then $\delta_j + \beta_j = \beta_{j-1}$ and F is a discrete Dirichlet process, with parameter $cH(\cdot)$.

Lemma A2 *If* $V_j = Y_j/(1-\sum_{s=0}^{j-1} Y_s)$, *then for any* $m > 0$ *the random variables* $V_0, V_1, \cdots, V_m$ *are independent and marginally distributed as* $\text{Beta}(\delta_j, \beta_j)$. *Furthermore, inverting the relation among the* $Y's$ *and the* $V's$, *gives the representation* $Y_0 = V_0$, $Y_1 = V_1(1 - V_0)$, $\ldots Y_m = V_m \prod_{s=0}^{m-1}(1 - V_s)$.

Lemma A3 *Let* $x_1, \ldots, x_n$, *with each* $x_s \in \{0, 1, \cdots j \cdots\}$, *be an iid sample from an unknown distribution function* F *on* $[0, \infty)$, *if the prior for* F *is chosen to be a discrete Beta-Stacy process with parameters* δ_j, β_j *and jumps at* $\{0, 1, \cdots j \cdots\}$, *then:*

(i) Given $x_1, \cdots, x_n$ *the posterior for* F *is also a discrete Beta-Stacy process with parameters* δ_j^*, β_j^* , *where* $\delta_j^* = \delta_j + r_j$, $\beta_j^* = \beta_j + m_j$, $m_j = \sum_{s>j} r_s$, *and* r_j *is the number of observations at the jump point* j*-th.*

(ii) The predictive probability $\mathbb{P}(x_{n+1} = j | x_1, \cdots x_n)$ *is given by*

$$\mathbb{E}[F(j) - F(j-1)|x_1, \ldots, x_n] = \frac{\delta_j^*}{\delta_j^* + \beta_j^*} \prod_{s=0}^{j-1} \frac{\beta_s^*}{\delta_s^* + \beta_s^*}. \tag{33}$$