# BAYESIAN TESTING OF A PARAMETRIC MODEL VERSUS NONPARAMETRIC ALTERNATIVES

JAMES O. BERGER AND ALESSANDRA GUGLIELMI

*ISDS, Duke University and CNR-IAMI, Milano*

11th January 1999

**Abstract**

Testing the fit of data to a parametric model can be done by embedding the parametric model in a nonparametric alternative and computing the Bayes factor of the parametric model to the nonparametric alternative. Doing so by specifying the nonparametric alternative via a Polya tree process is particularly attractive, from both theoretical and methodological perspectives. Default (non-subjective) versions of this analysis are developed herein, and illustrated on various examples.

*Keywords*: Testing fit; Nonparametric Bayes; Polya tree process; Bayesian robustness; Bayes factor; noninformative priors.

## 1. Introduction

For real i.i.d. observations, $X_1, \ldots, X_n$, we consider the problem of testing the fit of a parametric model $H_0 : f(x_i|\theta)$, where $f$ is assumed to be a density with respect to Lebesgue measure and $\theta \in \Theta$ is unknown. This will be done by formally testing $H_0$ against a nonparametric alternative, $H_1$. The approach taken is Bayesian, assigning prior distributions and computing the Bayes factor of $H_0$ to $H_1$, but we will primarily be concerned with doing so in a non-subjective fashion; in particular, for the parametric model we will just assume use of standard noninformative priors for $\theta$.

Consideration of this problem from the Bayesian perspective has a relatively recent history. In Florens, Richard, and Rolin (1996), a parametric model is compared to a nonparametric alternative by utilizing a Dirichlet process prior on alternative distributions. The Dirichlet process is chosen in such a way that the predictive distribution of a single observation is the same under the two hypotheses; this allows the alternative to be viewed as a generalization of the null model. Another natural approach to defining a nonparametric Bayesian alternative is to embed the parametric model directly into the nonparametric alternative. For instance, in Carota and Parmigiani (1996), the parametric model is embedded into a mixture of Dirichlet processes by assuming that, given $\theta$, the mean of the mixture Dirichlet process prior is equal to the corresponding null distribution for a single observation. Thus the Dirichlet process is also indexed by $\theta$, which is then assigned the same prior distribution as used for the parametric model.

As observed in the above papers, however, the use of Dirichlet processes is not really appropriate for testing an absolutely continuous parametric null model. The chief difficulty, as observed in Florens, Richard and Rolin (1996), is that the Bayes factor of $H_0$ to $H_1$ contains a penalty term that arises because the Dirichlet process 'expects' to see ties in the data. Indeed, as more and more distinct observations arrive (and they must be distinct since we are assuming that $f(x|\theta)$ is absolutely continuous), the penalty term will typically drive the Bayes factor to infinity, even when the null model is false. Also, Carota and Parmigiani (1996) show that, when the parametric model is absolutely continuous and no duplicate observations occur, the Bayes factor depends on the data only through the sample size, a property that is not very appealing.

Verdinelli and Wasserman (1998) approached the problem by considering a mixture of Gaussian processes to define the alternative, rather than a mixture of Dirichlet processes. In particular, they start with a random variable $U$ on the unit interval, whose log density is an infinite series of orthogonal polynomials with Gaussian coefficients, and then assume the i.i.d. observations $X_i$ have the same distribution as $F_\theta^{-1}(U)$, where $F_\theta$ is the cumulative distribution function corresponding to $f(x|\theta)$. Verdinelli and Wasserman (1998) utilize Bayes factors to perform the test, while Viele (1996) considers a similar model but uses a different goodness-of-fit tool.

In this paper, we consider a third mixture model for the alternative, namely a mixture of Polya tree processes. Our original motivation for considering this approach was twofold. First, Polya trees processes are a popular alternative to Gaussian processes for nonparametrically modeling continuous densities. Second, it seemed that a mixture of Polya trees would be a promising model for ensuring that prior distributions are appropriately 'calibrated' across hypotheses. This is of particular concern when noninformative prior distributions are to be used, as will be the case here for $\theta$ (cf., Berger and Pericchi, 1996, and Berger, Pericchi, and Varshavsky, 1998). Although $\theta$ will occur under both $H_0$ and $H_1$, it could have very different meanings under each and could, therefore, require different noninformative priors. We will argue in Section 3 that, with the Polya tree alternative, the usual noninformative priors for $\theta$ are appropriately calibrated, at least for the examples we consider. See, also, MacEachern (1998) for related ideas.

In the course of the investigation, we encountered a number of other attractive properties of the formulation for tackling the goodness-of-fit problem. (Some of these advantages were already recognized as inherent in any use of Polya tree processes; cf. Lavine, 1992, 1994). For convenience, these properties are summarized here, grouped by category.

*Attractive Features of the Bayesian Approach:* Any Bayesian approach to testing fit has certain appealing properties that are worth reviewing.

• Bayesian answers have a clear interpretability, as compared with the commonly used P-value. Delampady and Berger (1990) illustrate this issue in parametric testing of fit, demonstrating that P-values can dramatically overstate the actual evidence against a null model. For instance, in chi-squared testing of fit, it was shown that a P-value of 0.05 typically corresponds to at most 3 to 1 evidence against the null model.

• As it is not based on asymptotics, the Bayesian approach can handle small sample sizes. For instance, later examples consider testing fit to a normal distribution with unknown mean and variance, based on as little as 10 observations.

• There is the possibility of subjectively specifying the degree of departure from the null model that is anticipated. We will not pursue this aspect of the situation, however, as our focus is on non-subjective testing of fit.

*Attractive Features of the Polya Tree Process:*

• It is easy to embed the null parametric model into the alternative, by specifying it to be the mean of the Polya tree process (as with the Dirichlet process).

• With this embedding, it will be shown that usual noninformative priors for $\theta$ are appropriately calibrated across models, the motivating goal mentioned earlier.

• The Polya tree process has a rich structure, with numerous 'free parameters' remaining after specification of the mean. (In contrast, the Dirichlet process has only one free parameter after specifying the mean.) The Polya tree parameters include variance specifications and allow choice of the smoothness of the alternative densities.

• Polya tree processes have considerable promise as priors for higher dimensional distributions, and many of the techniques utilized here should readily adapt to higher dimensions.

*Computational Advantages of the Approach:* Computation of Bayes factors is typically considerably more challenging than computation of other Bayesian quantities, and nonparametric Bayesian analysis is considerably more difficult than parametric Bayesian analysis. It is thus a considerable surprise that computations under the structure adopted here are quite fast. The following are the reasons for the efficiency:

• A fixed partition for the Polya tree can be used (i.e., a partition not depending on $\theta$), and this has numerous computational advantages.

• The needed Polya tree predictive distributions corresponding to fixed $\theta$ have a finite exact expression, involving on the order of $n \log(n)$ quantities; other approaches typically deal with products (over $n$) of an infinite number of terms, and one then needs to find an appropriate truncation.

• Integrations over $\theta$ can sometimes be handled by ordinary MCMC, using a sample from the parametric model posterior.

*Robustness Advantages of the Approach:* Because of the richness of the Polya tree process and the computational speed of the analysis, it is possible to conduct robustness studies with respect to choice of the prior on the alternative. This is particularly important in Bayesian nonparametrics since very few features of a nonparametric prior will typically be specified subjectively, even if desired. Robustness studies of two types are conducted:

- A variety of choices are considered for the smoothness of the densities allowed in the alternative; this is controlled by the rate of decrease of the variances as one descends the Polya tree.

- The lower bound on the Bayes factor over choice of the overall scale for the prior variance is found; this allows for a degree of adaptability of the alternative to the data, useful for checking fit.

The paper is organized as follows. Section 2 gives the structure of the mixture of Polya tree processes that will be used to define $H_1$, and gives needed properties of the process. In Section 3, the Bayes factor of $H_0$ to $H_1$ is defined and the issue of calibrating the noninformative prior between the hypotheses is considered. Section 4 gives the computational algorithms used to compute the Bayes factor. Examples are presented in Section 5, and conclusions in Section 6.

It should be emphasized that this paper is primarily a *methodological* exploration of the use of mixtures of Polya trees in nonparametric testing of fit. The focus is on consideration and comparison of the myriad of possible choices that can be made in defining the Polya trees and needed priors.

# 2. Polya Trees and Mixtures of Polya Trees

## 2.1 Definition and Basic Properties

Polya tree priors are a special case of tailfree processes (see Freedman, 1963, and Fabius, 1964). Reference papers on Polya trees are Ferguson (1974), Mauldin, Sudderth and Williams (1992), and Lavine (1992, 1994), whose notation we will follow. Let $E = \{0, 1\}$, $E^0 = \emptyset$, $E^m$ be the $m$-fold product $E \times \cdots \times E$, and $E^* = \cup_0^{+\infty} E^m$. Let $\pi_0 = \mathbb{R}$ and $\Pi = \{\pi_m, m = 0, 1, \dots\}$, where $\pi_0, \pi_1, \dots$ is a sequence of partitions of $\mathbb{R}$ such that $\cup_0^{+\infty} \pi_m$ generates the Borel sets and every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\emptyset = \mathbb{R}$ and, for all $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two pieces into which $B_\varepsilon$ is split.

A random probability measure $\mathcal{P}$ on $\mathbb{R}$ has a Polya tree distribution with parameter $(\Pi, \mathcal{A})$ ($\mathcal{P} \sim PT(\Pi, \mathcal{A})$), if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\varepsilon, \varepsilon \in E^*\}$ and random variables $\mathcal{Y} = \{Y_\varepsilon, \varepsilon \in E^*\}$, such that

- all the random variables in $\mathcal{Y}$ are independent;
- for every $\varepsilon \in E^*$, $Y_\varepsilon$ has a Beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$ ;
- for every $m = 1, 2, \dots$ and every $\varepsilon \in E^*$,

$$\mathcal{P}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \prod_{j=1, \varepsilon_j = 0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \prod_{j=1, \varepsilon_j = 1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}).$$

For instance, for $m = 2$, $\mathcal{P}(B_{00}) = Y_\emptyset Y_0$, $\mathcal{P}(B_{01}) = Y_\emptyset(1 - Y_0)$, $\mathcal{P}(B_{10}) = (1 - Y_\emptyset) Y_1$ and $\mathcal{P}(B_{11}) = (1 - Y_\emptyset)(1 - Y_1)$. The $\alpha_\varepsilon$'s affect the Polya tree in two ways that are of interest here. First, the $\alpha_\varepsilon$'s determine the smoothness of the realizations of $\mathcal{P}$; second, they control how closely the distribution of $\mathcal{P}$ is concentrated around its mean (see Lavine, 1992).

A random probability measure is said to be a mixture over $\theta$ of Polya trees, with mixing distribution $\pi$ and parameters $\{\Pi_\theta, \mathcal{A}_\theta\}$, $\theta \in \Theta$, if the conditional distribution of $\mathcal{P}$, given $\theta$, is $PT(\Pi_\theta, \mathcal{A}_\theta)$. Slightly abusing notation, $\mathcal{P}_\theta$ will be used to denote a Polya tree random probability measure with distribution $PT(\Pi_\theta, \mathcal{A}_\theta)$.

### 2.2 Embedding a Parametric Model Into a Polya Tree

As discussed in the introduction, we will need to embed the parametric model $f(x|\theta)$ into the Polya tree process. Letting $\mu_\theta^f$ denote the probability measure corresponding to $f$, this can be accomplished by choosing a mixture of Polya trees, with parameters $(\Pi_\theta, \mathcal{A}_\theta)$, such that

$$E\mathcal{P}_\theta = \mu_\theta^f, \quad \forall \theta \in \Theta. \tag{1}$$

There are two "standard" Polya trees for which (1) holds, to be denoted by $\mathcal{P}_\theta^1$ and $\mathcal{P}_\theta^2$, respectively. $\mathcal{P}_\theta^1$ is the canonical construction of a Polya tree with fixed mean (see Lavine, 1992, section 2.3): if $F_\theta$ is the c.d.f. corresponding to $f(x|\theta)$, then choosing

$$\pi_{\theta,m} = \left\{ \left( F_\theta^{-1}\left(\frac{k}{2^m}\right), F_\theta^{-1}\left(\frac{k+1}{2^m}\right) \right], k = 0, 1, \ldots, 2^m - 1 \right\}, m = 1, 2, \ldots,$$

and $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ for any $\varepsilon \in E^*$, the $\alpha$'s not depending on $\theta$, results in a Polya tree whose mean is precisely $\mu_\theta^f$.

To define $\mathcal{P}_\theta^2$, first choose a fixed binary tree $\Pi$ of partitions, not depending on $\theta$, given by

$$\pi_m = \left\{ \left( \widetilde{F}^{-1}\left(\frac{k}{2^m}\right), \widetilde{F}^{-1}\left(\frac{k+1}{2^m}\right) \right], k = 0, 1, \ldots, 2^m - 1 \right\}, \tag{2}$$

$m = 1, 2, \ldots$, where $\widetilde{F}$ is a fixed distribution deemed to be representative of $H_0$. Then choose the $\alpha$'s, depending on $\theta$, so that (1) holds. Defining

$\underline{\varepsilon}_{m-1} = \varepsilon_1 \cdots \varepsilon_{m-1}$ and

$$c_{\varepsilon_1 \cdots \varepsilon_{m-1}}(\theta) = \frac{\mu_\theta^f(B_{\varepsilon_1 \cdots \varepsilon_{m-1} 1})}{\mu_\theta^f(B_{\varepsilon_1 \cdots \varepsilon_{m-1} 0})}, \quad m = 1, 2, \ldots ,$$

(for $m = 1$, $c_\emptyset(\theta) = \mu_\theta^f(B_1)/\mu_\theta^f(B_0)$), a natural class of such $\alpha$'s is given by

$$\begin{aligned} \alpha_{\underline{\varepsilon}_{m-1} 0}(\theta) &= \frac{d(\underline{\varepsilon}_{m-1})}{\sqrt{c_{\underline{\varepsilon}_{m-1}}(\theta)}} \\ \alpha_{\underline{\varepsilon}_{m-1} 1}(\theta) &= d(\underline{\varepsilon}_{m-1}) \sqrt{c_{\underline{\varepsilon}_{m-1}}(\theta)}. \end{aligned} \tag{3}$$

This class of $\alpha$'s not only satisfies (1), but also has the desirable symmetry property of being invariant to the choice of the labels '0' and '1'. In the following we shall use this second construction to define the Polya tree process. The reasons for this choice are primarily computational; having a fixed tree structure is an advantage at several stages of the computation. It remains only to specify $\widetilde{F}$ and the $d(\underline{\varepsilon})$.

Since $\widetilde{F}$ determines the partitions used for the Polya tree, it is intuitively clear that this should be chosen to provide a good partition of the actual data. We shall choose $F = F_{\hat{\theta}}$, where $\hat{\theta}$ is the m.l.e. for $\theta$ with respect to the model $f(x|\theta)$. Another choice for $F$ that is very attractive from a computational perspective is the empirical distribution function, in that this can be shown to minimize the number of levels needed for the Polya tree computation. This choice has uncertain theoretical properties, however, because of the extreme data-dependence of the partition, so we will not consider it herein.

As to choice of the $d(\underline{\varepsilon}_m)$, we consider and compare a variety of possibilities of the form

$$d(\underline{\varepsilon}_m) = h^{-1} \cdot \rho(m), \tag{4}$$

where $h$ refers to an overall scale factor and $m$ denotes the level of the Polya tree corresponding to $\underline{\varepsilon}_m$. The function $\rho(\cdot)$ controls the smoothness of the distributions upon which the Polya tree process concentrates its mass (Lavine, 1992). For instance, it is remarked in Ferguson (1974) that, if $d(\underline{\varepsilon}_{m-1})$ is proportional to $m^2$, then the realizations of the Polya tree process are guaranteed to be absolutely continuous (with probability one); see Draghici and Ramamoorthi (1998) for more general results in this direction. Choices of $\rho(m)$ that have exponential growth are motivated by consistency considerations (Barron, Schervish and Wasserman, 1997). In addition to these 'standard' growth rates, a simple heuristic argument given in the Appendix suggests choosing

$$d(\underline{\varepsilon}_{m-1}) \propto \frac{1}{h} \cdot \frac{f^2(x^*_{\underline{\varepsilon}_m}|\theta)}{(\mu^f_\theta(B_{\underline{\varepsilon}_{m-1}}))^2} \ , \tag{5}$$

where $x^*_{\underline{\varepsilon}_m}$ denotes the splitting point of $B_{\underline{\varepsilon}_{m-1}}$ into $B_{\underline{\varepsilon}_{m-1}0}$ and $B_{\underline{\varepsilon}_{m-1}1}$.

The scale, $h$, which in some sense controls the overall variance of the Polya tree about its mean (the parametric model), is very difficult to specify in nonparametric scenarios. Hence we deal with this parameter by a robustness analysis that will be further discussed in Section 5.

### 2.3 The Polya Tree Marginal Density

For given $\theta$, the marginal density of a sample, $X_1, \dots, X_n$, from a Polya tree can be easily computed, when it exists. Indeed, for $\mathcal{P}^2_\theta$, it follows from Lavine (1992) that this density is given by

$$m_1(x_1, \dots, x_n|\theta) = f(\underline{x}|\theta)\psi(\theta), \tag{6}$$

where $\underline{x} = (x_1, \dots, x_n)$, $f(\underline{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$, and

$$\psi(\theta) = \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha'_{\underline{\varepsilon}_m(x_j)}(\theta)(\alpha_{\underline{\varepsilon}_{m-1}0(x_j)}(\theta) + \alpha_{\underline{\varepsilon}_{m-1}1(x_j)}(\theta))}{\alpha_{\underline{\varepsilon}_m(x_j)}(\theta)(\alpha'_{\underline{\varepsilon}_{m-1}0(x_j)}(\theta) + \alpha'_{\underline{\varepsilon}_{m-1}1(x_j)}(\theta))} \ ; \tag{7}$$

here $\underline{\varepsilon}_m(x_j)$ is the index $\varepsilon_1 \cdots \varepsilon_m$ such that $x_j$ belong to $B_{\varepsilon_1 \cdots \varepsilon_m}$, for each level $m$, and $\alpha'_{\underline{\varepsilon}_m(x_j)}(\theta)$ is equal to $\alpha_{\underline{\varepsilon}_m(x_j)}(\theta)$ plus the number of observations among $\{x_1, \ldots, x_{j-1}\}$ that belong to $B_{\varepsilon_1 \cdots \varepsilon_m(x_j)}$. For each $x_j$, the product in (7) is up to the smallest level, $m^*(x_j)$, such that no $x_i$, $i < j$, belongs to $B_{\underline{\varepsilon}_m(x_j)}$. Thus, note that, if $M$ is the first level such that each interval in $\pi_M$ contains at most one observation, then $m^*(x_j) \leq M$ for all $j = 1, \ldots, n$. Furthermore, neither $M$ nor the $m^*(x_j)$ will depend on $\theta$, since the $\pi_m$ do not depend on $\theta$. This accounts for one of the computational simplifications in use of $\mathcal{P}^2_\theta$.

# 3. The Bayes Factor of $H_0$ to $H_1$

Using the embedding from Section 2.2, we desire to test $H_0 : f(x|\theta)$, $\theta \in \Theta$, versus $H_1 : \mathcal{P}^2_\theta$, $\theta \in \Theta$, utilizing a prior density, $\pi(\theta)$, that will be taken to be noninformative in examples. The test will be performed by computing the Bayes factor of $H_0$ to $H_1$, given by

$$BF(\underline{x}) = \frac{m_0(x_1, \ldots, x_n)}{m_1(x_1, \ldots, x_n)}, \tag{8}$$

where

$$m_0(x_1, \ldots, x_n) = \int_\Theta f(\underline{x}|\theta)\pi(\theta)d\theta$$

is the predictive (or marginal) density under $H_0$, and

$$m_1(x_1, \ldots, x_n) = \int_\Theta m_1(x_1, \ldots, x_n|\theta)\pi(\theta)d\theta$$

is the predictive density under $H_1$, with $m_1(x_1, \ldots, x_n|\theta)$ as defined in (6).

In the above expression, we have committed the all-too-common sin of assuming that $\theta$ can be assigned the same prior distribution, $\pi(\theta)$, under $H_0$ and $H_1$. As mentioned in the introduction, this actually requires extrinsic justification. (Using the same letter for parameters in two different models does not make them the same parameter!)

A partial justification arises because of the embedding discussed in Section 2.2. Indeed, because of this embedding, it is easy to see that, for any proper $\pi(\theta)$, the predictive distribution of a single (hypothetical) observation, $x_1$, will be the same for both $H_0$ and $H_1$ (i.e., $m_0(x_1) = m_1(x_1)$). However, as argued in Berger and Pericchi (1996) and Berger, Pericchi, and Varshavsky (1998), the ideal is to have $m_0(x_1, \ldots, x_m) = m_1(x_1, \ldots, x_m)$ for any (hypothetical) sample of size $m$, where (usually) $m$ is the dimension of $\theta$. The simplest intuition for this result is that a sample of size $m$ is needed to even identify $\theta$ in each model, after which there are no 'degrees of freedom' left for model comparison. Hence a sample of size $m$ should always yield a Bayes factor of 1 (implying that the two models are equally supported by the data) and this will indeed be the case if $m_0(x_1, \ldots, x_m) = m_1(x_1, \ldots, x_m)$. This condition is termed 'predictive matching' in Berger and Pericchi (1996).

The examples we consider later have $\theta = (\mu, \sigma)$, with $(\mu, \sigma)$ being a location-scale parameter for $f(x|\theta)$, i.e., the density is of the form $\sigma^{-1}p((x - \mu)/\sigma)$. The number of observations needed to 'identify' the parameters from such a distribution is $m = 2$. If the right Haar prior, $\pi(\mu, \sigma) = 1/\sigma$ is used, it is known (see Berger, Pericchi and Varshavsky, 1998) that $m(x_1, x_2) = 1/(2|x_1 - x_2|)$ for any two distinct observations and for any absolutely continuous $p$. Thus, if it can be shown that the distribution under $H_1$ is also an absolutely continuous location-scale distribution, it will follow that $m_0(x_1, x_2) = m_1(x_1, x_2)$, as desired, providing $\pi(\mu, \sigma) = 1/\sigma$ is also used for $H_1$.

In this situation, it is straightforward to show that $\mathcal{P}^1_\theta$ has a location-scale distribution and that, if the $\alpha_{\underline{\varepsilon}_m}$ increase at least as fast as $m^2$ (the level corresponding to $\varepsilon_m$), then $\mathcal{P}_\theta^1$ is absolutely continuous. Hence, it follows that $H_0 : f(x|\theta)$ and $H_1 : \mathcal{P}_\theta^1$ would be predictively matched (assuming the right Haar prior is used). Indeed, this predictive matching can be shown to hold whenever $f(x|\theta)$ possesses a suitable invariance structure, using the more general results in Berger, Pericchi, and Varshavsky (1998).

Unfortunately, $\mathcal{P}_\theta^2$ will not, in general, inherit the invariance structure of $f(x|\theta)$, since the partition defining this Polya tree is fixed (i.e., does not depend on $\theta$). Intuitively, however, the construction of $\mathcal{P}_\theta^2$ is such that one would expect approximate predictive matching for the right Haar prior, and this will be shown to be the case in the location-scale examples considered later. If one were concerned that approximate predictive matching would not obtain, one could use, say, the median intrinsic Bayes factor of Berger and Pericchi (1998), defined as

$$(9) \qquad BF^I(\underline{x}) = \frac{m_0(x_1, \ldots, x_n)}{m_1(x_1, \ldots, x_n)} \cdot \text{Median} \left\{ \frac{m_1(x_1(k), x_2(k))}{m_0(x_1(k), x_2(k))} \right\},$$

where the median is taken over all distinct pairs of observations $(x_1(k), x_2(k))$ (or a random subset of such pairs of size $n$). The idea here is simply to multiply the Bayes factor by a correction term that would compensate for the lack of predictive matching between $H_0$ and $H_1$. See Berger and Pericchi (1998) for more discussion.

## 4. Computing the Bayes Factor

Two different Monte Carlo techniques are considered for computing $BF(\underline{x})$ in (8). The first is the standard (but often inefficient) approach based on direct sampling from the null posterior. Note, from (6), (7) and (8), that the

Bayes factor can be written as

$$BF(\underline{x}) = \left( \int_{\Theta} \psi(\theta)\pi(\theta|\underline{x})d\theta \right)^{-1},$$

where $\pi(\theta|\underline{x}) = f(\underline{x}|\theta)\pi(\theta)/m_0(\underline{x})$. This suggests generating a sample, $\theta_1, \ldots, \theta_L$, from $\pi(\theta|\underline{x})$, and approximating the Bayes factor by

$$\widehat{BF}_1(\underline{x}) = \frac{L}{\sum_{j=1}^{L} \psi(\theta_j)} \ . \tag{10}$$

For the normal examples considered in Section 5, it can be shown that $\int \psi^r(\theta)\pi(\theta|\underline{x})d\theta < \infty$ only for $r < 2$. Hence $\widehat{BF}_1$ will converge to $BF$ as $L \to \infty$, but $\widehat{BF}_1$ does not have a variance and the convergence might be slow. Nevertheless, the simplicity of this approach, especially when a sample from $\pi(\theta|\underline{x})$ is available, is appealing. Note, in particular, that Bayesian analysis of an entertained model usually involves generating a (possibly dependent) sample from $\pi(\theta|\underline{x})$ via an MCMC algorithm, and that testing of fit then requires only the additional computation of (10).

The above simulation can be inefficient because $\pi(\theta|\underline{x})$ can have tails that are too sharp. The convergence speed can often be improved by using an importance density, $q(\theta)$, that approximates $\pi(\theta|\underline{x})$ but has flatter tails. Then, if $\theta_1, \ldots, \theta_L$ is an i.i.d. sample from $q$, the importance sampling approximation to the Bayes factor is

$$\widehat{BF}_2(\underline{x}) = \frac{\sum_{j=1}^{L} f(\underline{x}|\theta_j)\pi(\theta_j)/q(\theta_j)}{\sum_{j=1}^{L} \psi(\theta_j) f(\underline{x}|\theta_j)\pi(\theta_j)/q(\theta_j)}. \tag{11}$$

The value of having a fixed partition for $\mathcal{P}_\theta^2$ is apparent when considering either of the above computational schemes. Then the tree structure that defines the products in (7) has a fixed structure that can be computed and tabled before the simulation loop. In contrast, with $\mathcal{P}_\theta^1$ the entire structure of the tree changes with each $\theta_j$ and must be recomputed each time inside the loop. The need for efficiency here is particularly pronounced because of the fact that, in Section 5, we will simultaneously need to compute hundreds of Bayes factors, as the scale factor $h$ in (4) and (5) varies. This need is also a reason why we do not consider more elaborate MCMC computational schemes. Either of the above simulation schemes can be readily adapted to efficient simultaneous computation of hundreds of Bayes factors; the same posterior or importance samples can be used for all the Bayes factors, and the expressions being computed only change slightly with $h$.

The lack of a variance makes it somewhat difficult to obtain an estimate of the accuracy of $\widehat{BF}_1$, but the accuracy of $\widehat{BF}_2$ is easy to determine. Indeed, the numerical standard error of the approximation, $\hat{\sigma}_2$, can be estimated using the expressions in, say, Geweke (1989). It is most convenient to report the relative error, $\hat{\sigma}_2/\widehat{BF}_2$, and this will be done in the examples.

The examples in the next section involve testing the normal model with likelihood $f(\underline{x}|\mu,\sigma^2)$. The importance density, $q(\mu,\sigma^2)$, that will be used to approximate this likelihood is

$$(12)\qquad q(\mu,\sigma^2) = \frac{n}{\sqrt{2}\pi c\hat{\sigma}\sigma^2}\left[1+\frac{n}{2c\hat{\sigma}^2}(\mu-\hat{\mu})^2+\frac{n}{4c}(\log\frac{\sigma^2}{\hat{\sigma}^2})^2\right]^{-3},$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the m.l.e.'s for $\mu$ and $\sigma^2$. This arises from considering the variables $\mu$ and $\xi = \log(\sigma^2)$, both of which are location variables. A reasonable importance density for location variables is the bivariate Student-t density with 4 degrees of freedom, mean equal to the m.l.e., $(\hat{\mu}, \log(\hat{\sigma}^2))$, and covariance matrix given by a constant $c$ times $\hat{I}^{-1}$, where $\hat{I}$ is the observed Fisher information matrix in this parameterization. Transforming back to $(\mu,\sigma^2)$ results in the importance function in (12). Note that, in simulations, it is easiest to generate $(\mu,\xi)$ directly from the t-density, and then simply set $\sigma^2 = \exp(\xi)$. Observe that $c$ controls the spread of the importance density; in examples we try a variety of $c$ to optimize the speed of convergence.

# 5. Examples

We consider two examples involving tests of the null hypothesis that the data come from a normal distribution with mean $\mu$ and variance $\sigma^2$. The alternative considered in both examples is $\mathcal{P}^2_\theta$, where the partitions are as in (2) with $\widetilde{F} = F_{\hat{\theta}}$, $\hat{\theta}$ being the maximum likelihood estimator of $\theta = (\mu, \sigma^2)$. The $\alpha$'s are chosen as in (3), and the Bayes factors are determined from (8) with $\pi(\mu, \sigma) = 1/\sigma$, the usual noninformative prior.

The remaining features of $\mathcal{P}^2_\theta$ that need to be specified are the $d(\underline{\varepsilon}_{m-1})$ in (3). We will consider a variety of choices in the examples, specifically the choices in (4) and in (5), the former with $\rho(m)$ proportional to $m^2$, $m^3$, $2^m$, $4^m$, or $8^m$. It remains only to specify the overall scale factor $h$, which controls the overall variance of the Polya tree about its mean (the parametric model).

The scale $h$ is a very influential variable in testing, and is quite difficult to specify in nonparametric scenarios. Hence we deal with $h$ by a robustness analysis, presenting the Bayes factor as a function of $h$. For testing a precise null against a vaguer alternative (such as here), the possible behaviors of the Bayes factor as a function of $h$ are as in Figure 2. As $h \to 0$, the Polya tree process will collapse to its mean (the parametric model), and the Bayes factor will converge to 1 (or 0 on the log scale, as in Figure 2). As $h \to \infty$, the Polya tree process will (typically) become so vague that it gives little mass to any distributions reflective of the actual data; the Bayes factor of $H_0$ to $H_1$ will then be quite large. Between these extremes, the Bayes factor will sometimes be increasing in $h$, as in Figure 2(a), but often will first decrease and then increase, as in the other cases in Figure 2.

A particularly interesting point is the minimum of the Bayes factor over $h$, since this indicates the prior 'scale' which is most favorable to the alternative. If this minimum is not small, then there is no reason to doubt $H_0$. Of course, even if this minimum is small one should not summarily reject $H_0$, since the minimum is achieved by searching for the most favorable prior for $H_1$ for the given data, which clearly results in a bias against $H_0$; but, at least, it is useful to know that there are alternatives that better explain the data. (See Bayarri and Berger, 1999, for further discussion of this issue.) This minimum is also of use in judging which of the 'rates of increase' in $d(\underline{\varepsilon}_m)$ are superior in terms of fitting the data, with a smaller minimum indicating a better fit for the corresponding alternative.

**Example 1.** For comparison with the results in Verdinelli and Wasserman (1998), we first consider the 100 stress-rupture lifetimes of Kevlar pressure vessels from Andrews and Herzberg (1985), p.183, desiring to test the null hypothesis of normality. The importance sampling scheme, discussed in Section 4, was utilized for the computations. The maximum number of levels required for the Polya tree was $M = 13$ but most of the $m^*(x_j)$ in (7) were less than 8. In computing $\widehat{BF}_2$ in (11), any value of $c$ between 10 and 100 worked well for the importance function; $c = 50$ was used for the computations reported below. Finally, $L = 10,000$ turned out to be sufficient to keep the relative Monte Carlo error, $\hat{\sigma}_2/\widehat{BF}_2$, less than 10%, and was utilized in the computations.

Figure 1 gives plots of $\log_{10} \widehat{BF}_2$, as a function of $h$, for various choices of $d(\underline{\varepsilon}_m)$ of the form (4) and (5), with the latter choice denoted therein by $d(eps)$ . Suitable proportionality constants were chosen for each form so that the minimums of the resulting Bayes factors were clearly revealed. These minimums are given in Table 1 for the various choices of $d(\underline{\varepsilon}_m)$, along with the minimizing $\hat{h}$ and the corresponding Monte Carlo relative error. These minimums are extremely small, indicating overwhelming evidence against $H_0$. Figure 1 also indicates that this conclusion would be reached for any reasonable $h$. We thus have compelling evidence against $H_0$.

It is of interest that the form of $d(\underline{\varepsilon})$ that gives the smallest minimum is that given in (5). This suggests that this choice allows $\mathcal{P}_\theta^2$ to best fit the data. Since the rate of increase of this choice is essentially exponential in $m$, it also yields a Polya tree that is supported on smooth densities, which would appear to avoid the danger of overfitting. Hence it appears that this is a very attractive choice for $d(\underline{\varepsilon})$. In this regard, it is also interesting to note that the Bayes factor in this example that was found in Verdinelli and Wasserman (1998) was 0.1. Apparently the prior on the alternative that was used therein is not as adaptable to the data as those considered here.

| $d(\underline{\varepsilon}_{m-1})$ | $\min_h(\log_{10}\widehat{BF}_2)$ | $\hat{h}$ | $\hat{\sigma}_2/\widehat{BF}_2$ |
|---|---|---|---|
| $10^6/h * f_\theta^2/\mu_\theta^2$ | -9.705073 | 6 | 5.3% |
| $\frac{5}{h}2^m$ | -8.480833 | 14 | 5.4% |
| $\frac{10}{h}m^2$ | -8.305651 | 24 | 5.7% |
| $\frac{1}{h}4^m$ | -8.294694 | 20 | 6.1% |
| $\frac{2}{h}m^3$ | -8.114254 | 14 | 6.2% |
| $\frac{1}{10\cdot h}8^m$ | -7.683233 | 13 | 6.6% |

Table 1. Minimum log Bayes factors for the Kevlar pressure vessels data.

Computation of $\widehat{BF}_1$ in (10) was also attempted in this example. The convergence rates were quite variable, however. For instance, when $d(\underline{\varepsilon}_{m-1}) = (5/14) * 2^m$, reasonable convergence was not obtained until $L = 10,000,000$. This indicates that this algorithm should indeed be used cautiously.

Finally, we investigated the calibration or matching issue discussed in Section 3, by computing $BF^I$ in (9) for many of the cases considered above. There was very little qualitative difference in the answers. Indeed, the ratio $BF^I/BF$ varied between 0.68 and 1.03, and was within 5% of 1 about 65%

of the time. Furthermore, $BF$ seems to err on the high side (if at all), which may help to compensate for the bias against $H_0$ induced by focusing on the minimum Bayes factor over $h$. Although this is only one example, it is a rather extreme case (as indicated by the extremely small Bayes factors), lending some assurance that the calibration issue can be ignored in use of $\mathcal{P}_\theta^2$ and that the more involved computation involving $\mathcal{P}_\theta^1$ can be avoided.

**Example 2.** To investigate if our nonparametric alternative might routinely overfit the data, we also analyzed data simulated from a normal distribution. For simulated samples of size $n = 100$, we found that the Bayes factor was virtually always quite large (even the minimums over $h$) , properly indicating that the null hypothesis of normality is true. We thus tried smaller sample sizes, for which the uncertainty would be greater. Of most interest were the results for samples of size $n = 10$. Ten such random samples were generated, and $\widehat{BF}_2$ was computed, as above, for each case. Since the various choices for the form of $d(\underline{\varepsilon}_m)$ gave similar results, we report only the results for the choice proportional to $m^2$.

For seven of the ten simulated data sets, the plot of the Bayes factor as a function of $h$ was as in Figure 2(a); the Bayes factor was always bigger than 1, indicating that there is no evidence against normality. For three of the data sets, the Bayes factor fell below 1 for some values of $h$, and these plots are given in Figure 2(b,c,d). The corresponding minimums were -1.04, -0.015 and -0.196, corresponding to Bayes factors of 0.091, 0.97, and 0.64, respectively. Clearly only the first of these would appear to suggest substantial evidence against normality. The actual (rounded and ordered) data set in this case was $\{-0.89, -0.88, -0.72, -0.66, 0.30, 0.35, 0.36, 0.73, 0.89, 2.60\}$. This certainly has the appearance of non-normal data; indeed, if one computes the sample mean and standard deviation of the first nine observations, the apparent 'outlying' tenth observation is almost four standard deviations away from the sample mean. Thus the fact that the Bayesian test indicates evidence against normality would hardly seem to be an egregious error for this data set.

## 6. Conclusions

In practice, the extreme flexibility of nonparametric Bayesian analysis, in terms of allowed choice of the nonparametric prior, is a disadvantage; subjective specification is extremely difficult, and it is often not even clear which features of the nonparametric prior (here a mixture of Polya tree processes) most affect the answer.

Testing of fit is a context in which it is possible to squarely address this issue and develop an essentially default Bayesian procedure. First, the nature of the problem of testing the fit of a parametric model essentially requires that the parametric model be chosen as the mean of the Polya tree process. Also, the (approximate) predictive matching of the ensuing process with the original parametric model allows a noninformative prior to be used for the unknown parameters. The remaining features of the mixture Polya tree process (the $d(\underline{\varepsilon})$) are not so easy to specify, and we studied a variety of choices. A tentative recommendation that can be given is to use the choice in (5), since this seemed to yield the most sensitivity to fit (and ensures that the nonparametric process gives mass only to reasonably smooth distributions).

Thus the only feature of the process that requires user specification is the scaling constant $h$, that affects the concentration of the nonparametric prior about the parametric family. One could argue for reporting the minimum Bayes factor over $h$ as an objective summary of fit. Indeed, this could be called the 'empirical Bayes BF,' since it arises from maximizing the denominator of (8) over $h$, i.e., results in use of the Type II maximum likelihood prior. While this is certainly a number to be reported and considered seriously (as it indicates how much better a fit can be obtained), it is clearly biased against the null model. Therefore, we recommend graphing the Bayes factor as a function of $h$, and calling this graph the 'objective summary' of the situation.

We found that use of the computationally simple $\mathcal{P}_\theta^2$, together with (8), is satisfactory in terms of calibration or predictive matching of the prior (i.e., it is not necessary to resort to use of the more involved (9) or $\mathcal{P}_\theta^1$). On the computational side, however, matters did not work out quite so well; the particularly simple computation via (10) can be slow to converge, so that the more involved importance sampling computation in (11) may be required.

Thus our final recommendation is to use the Bayes factor defined by (5) and (8), with the computations performed via either (10) (if it converges) or (11).

## Acknowledgments

This work was supported, in part, by the National Science Foundation of the U.S.A. under Grants DMS-9303556 and DMS-9802261.

## Appendix: Motivation for (5)

If $\alpha_{\underline{\varepsilon}_{m-1}0}$ and $\alpha_{\underline{\varepsilon}_{m-1}1}$ are given as in (3), and if the parametric density (as well the supposed alternative densities) can be considered to be almost constant on $B_{\underline{\varepsilon}_{m-1}} = (a, b]$ (which is a reasonable assumption if $m$ is large enough), then the variance of $Y_{\underline{\varepsilon}_m}$ is approximately equal to $(x^*_{\underline{\varepsilon}_m} - a)(b - x^*_{\underline{\varepsilon}_m})/[(b - a)^2(1 + d(\underline{\varepsilon}_{m-1})/\sqrt{c_{\underline{\varepsilon}_{m-1}}} + d(\underline{\varepsilon}_{m-1})\sqrt{c_{\underline{\varepsilon}_{m-1}}})]$. Ordinarily, this will be roughly proportional to $1/d(\underline{\varepsilon}_{m-1})$. On the other hand, the variation in plausible densities over $B_{\underline{\varepsilon}_{m-1}}$ will typically be proportional to $(b-a)$, which we can then equate with the square root of the variance of $Y_{\underline{\varepsilon}_m}$, which is proportional to $1/\sqrt{d(\underline{\varepsilon}_{m-1})}$. The implication is that it is reasonable to choose $d(\underline{\varepsilon}_{m-1})$ proportional to $(b-a)^{-2}$. This, in turn, can be approximated by $f^2(x^*_{\underline{\varepsilon}_m}|\theta)/(\mu^f_\theta(B_{\underline{\varepsilon}_{m-1}}))^2$, resulting in the suggested choice in (5).

## References


Andrews, D.F. and Herzberg, A.M. (1985), *Data: a collection of problems from many fields for the student and research worker*, New York: Springer-Verlag.

Barron, A., Schervish, M.J. and Wasserman, L. (1997), “The consistency of posterior distributions in nonparametric problems”, Technical Report 651, Department of Statistics, Carnegie-Mellon University.

Bayarri, M.J. and Berger, J.O. (1999), “Quantifying Surprise in the Data and Model Verification,” in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press (in press).

Berger, J.O. and Pericchi, L.R. (1996), “On the justification of default and intrinsic Bayes factors”, in *Modelling and prediction*, eds. J.C. Lee et. al., New York: Springer-Verlag, 276–293.

Berger, J.O. and Pericchi, L.R. (1998), “Accurate and Stable Bayesian Model Selection: the Median Intrinsic Bayes Factor,” *Sankhyā* , **B60**, 1–16.

Berger, J.O., Pericchi, L.R. and Varshavsky, J.A. (1998), "Bayes factors and marginal distributions in invariant situations", *Sankhyā* , **A60**, 307–321.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, Chichester: Wiley.

Carota, C. and Parmigiani, G. (1996), "On Bayes factors for nonparametric alternatives", in *Bayesian Statistics 5*, eds. J. M. Bernardo J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, 507–511.

Delampady, M. and Berger, J.O. (1990), "Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit", *The Annals of Statistics*, **18**, 1295–1316.

Draghici, L. and Ramamoorthi, R.V. (1998), "A note on the absolute continuity and singularity of Polya tree priors and posteriors", Technical Report, Michigan State University.

Fabius, J. (1964), "Asymptotic behaviour of Bayes' estimate", *Annals of Mathematical Statistics*, **35**, 846–856.

Ferguson, T.S. (1974), "Prior distributions on spaces of probability measures", *The Annals of Statistics*, **2**, 615–629.

Florens, J.P., Richard, J.F. and Rolin, J.M. (1996), "Bayesian encompassing specifications tests of a parametric model against a nonparametric alternative", Technical Report 96.08, Institut de Statistique, Université Catholique de Louvain, Belgium.

Freedman, D. (1963), “On the asymptotic behaviour of Bayes’ estimate in the discrete case”, *Annals of Mathematical Statistics*, **34**, 1194–1216.

Geweke, J. (1989), “Bayesian inference in econometric models using Monte Carlo integration”, *Econometrica*, **57**, 1317–1339.

Lavine, M. (1992), “Some aspects of Polya tree distributions for statistical modelling”, *The Annals of Statistics*, **20**, 1222–1235.

Lavine, M. (1994), “More aspects of Polya tree distributions for statistical modelling”, *The Annals of Statistics*, **22**, 1161–1176.

MacEachern, S.N. (1998), “Estimation of Bayes factors in mixture of Dirichlet process models”, Technical Report, Ohio State University.

Mauldin, R.D., Sudderth, W.D. and Williams, S.C. (1992), “Polya trees and random distributions”, *The Annals of Statistics*, **20**, 1203–1221.

Verdinelli, I. and Wasserman, L. (1998), “Bayesian goodness of fit testing using infinite dimensional exponential families”, *The Annals of Statistics*, **20**, 1203–1221.

Viele, K. (1996), “A nonparametric Bayesian method for evaluating fit in hierarchical curve models, Technical Report 96.358, Department of Statistics, University of Kentucky.

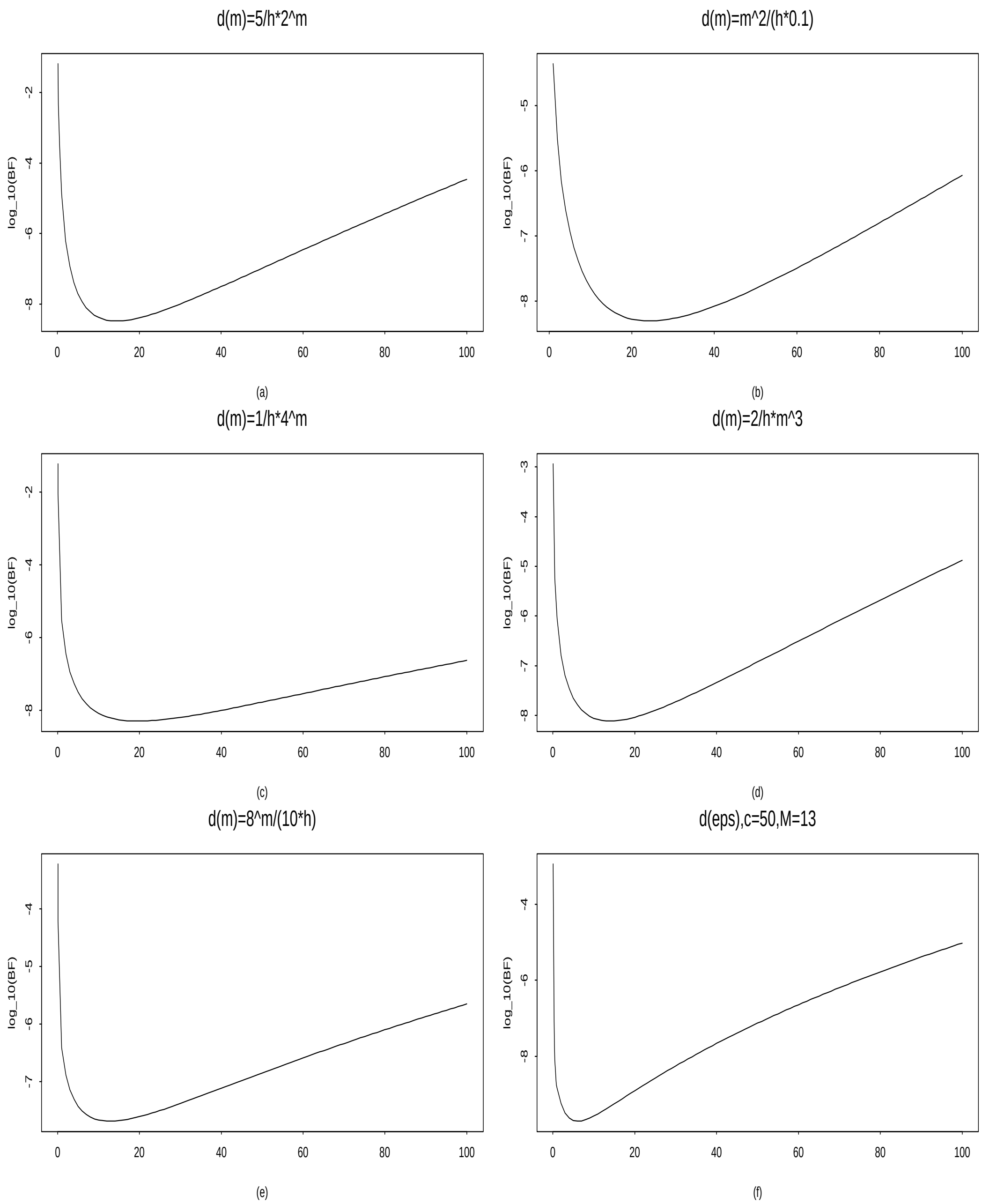
d(m)=5/h*2^m
log_10(BF)
(a)
d(m)=m^2/(h*0.1)
log_10(BF)
(b)
d(m)=1/h*4^m
log_10(BF)
(c)
d(m)=2/h*m^3
log_10(BF)
(d)
d(m)=8^m/(10*h)
log_10(BF)
(e)
d(eps),c=50,M=13
log_10(BF)
(f)

Figure 1

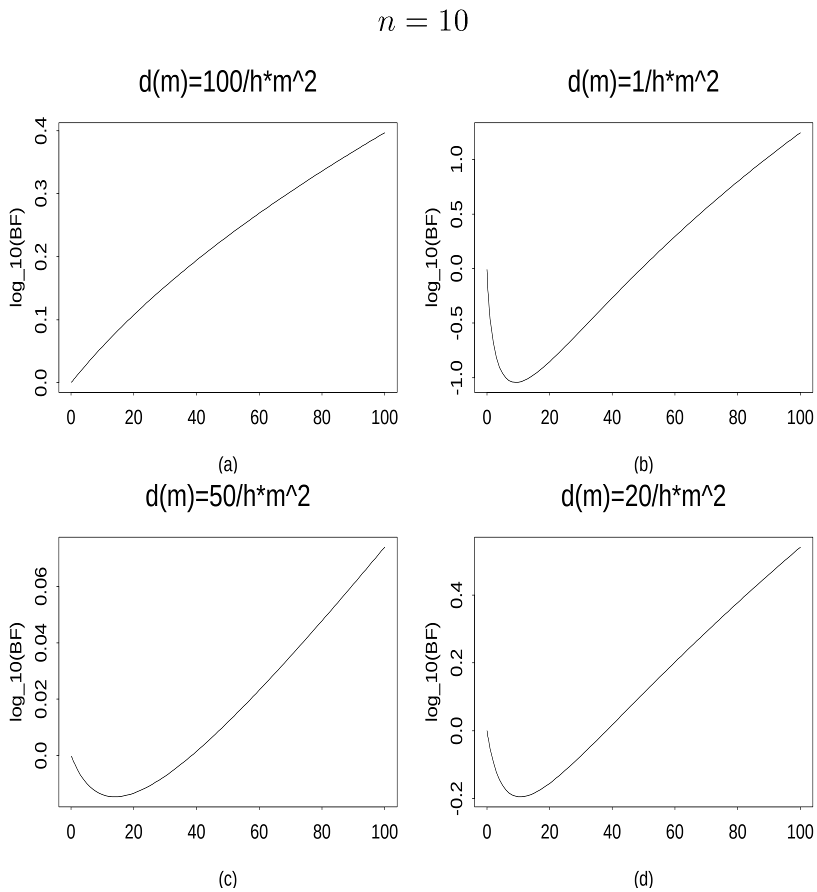
$n = 10$
d(m)=100/h*m^2
log_10(BF)
0.0
0.1
0.2
0.3
0.4
0 20 40 60 80 100
(a)
d(m)=1/h*m^2
log_10(BF)
-1.0
-0.5
0.0
0.5
1.0
0 20 40 60 80 100
(b)
d(m)=50/h*m^2
log_10(BF)
0.0
0.02
0.04
0.06
0 20 40 60 80 100
(c)
d(m)=20/h*m^2
log_10(BF)
-0.2
0.0
0.2
0.4
0 20 40 60 80 100
(d)

Figure 2