Statistica Sinica **15**(2005), 359-380

CALIBRATING BAYES FACTOR UNDER PRIOR PREDICTIVE DISTRIBUTIONS

Gonzalo García-Donato and Ming-Hui Chen

Universidad de Castilla-La Mancha and University of Connecticut

Abstract: The Bayes factor is a popular criterion in Bayesian model selection. Due to the lack of symmetry of the prior predictive distribution of Bayes factor across models, the scale of evidence in favor of one model against another constructed based solely on the observed value of the Bayes factor is thus inappropriate. To overcome this problem, a novel calibrating value of the Bayes factor based on the prior predictive distributions and the decision rule based on this calibrating value for selecting the model are proposed. We further show that the proposed decision rule based on the calibration distribution is equivalent to the surprise-based decision. That is, we choose the model for which the observed Bayes factor is less surprising. Moreover, we demonstrate that the decision rule based on the calibrating value is closely related to the classical rejection region for a standard hypothesis testing problem. An efficient Monte Carlo method is proposed for computing the calibrating value. In addition, we carefully examine the robustness of the decision rule based on the calibration distribution to the choice of imprecise priors under both nested and non-nested models. A data set is used to further illustrate the proposed methodology and several important extensions are also discussed.

Key words and phrases: Calibrating value, critical value, hypothesis testing, imprecise prior, L measure, model selection, Monte Carlo, posterior model probability, pseudo-Bayes factor, P-value.

1. Introduction

Let $\boldsymbol{y}$ denote the data and let M_1 and M_2 be two competing models. Then the Bayes factor, $\mathrm{B}(\boldsymbol{y})$, is one of the most popular criteria in selecting which of models M_1 and M_2 fit the data better. The Bayes factor can be interpreted as the strength of the evidence favoring one of the models for the given data. In fact, the Bayes factor is strongly related to the posterior probabilities of models M_1 and M_2, and hence its interpretation is straightforward. Jeffreys (1961) and Kass and Raftery (1995) have proposed the rules to determine the strength of evidence which can be associated with the observed value of the Bayes factor.

On the other hand, before the data are taken, the Bayes factor, $\mathrm{B}(\boldsymbol{Y})$, is a random variable, and its distribution follows that of $\boldsymbol{Y}$. Then the *sampling* properties of the Bayes factor can be examined under each of the two models under consideration. Once the data $\boldsymbol{Y} = \boldsymbol{y}$ are obtained, these properties can be used to *measure* the agreement between each of the two models and the data. According to this measure a decision can be made about the goodness-of-fit of each of the models. Thus, it is not appropriate to treat the Bayes factor as fixed once the data is observed due to uncertainty in the random sample. In particular, when the amount of uncertainty is large, the data, which are observed at the different time points but from the same distribution model, could lead to very different observed values of the Bayes factor. In addition, the decision rule that compares the observed Bayes factor to a predetermined value, called a "critical value", independent of the sampling distribution of the Bayes factor, can be misleading since the sampling distribution depends on the models being compared and the priors involved in deriving the posterior distributions. Therefore, it is important to calibrate the Bayes factor to take account of the randomness of the data. In fact, the idea of calibration is quite natural and appealing, since it is analogous to the classical testing hypotheses, in which the decision rule is established based on the sampling distribution of the testing statistic. The sampling arguments in selecting the model have been extensively used in the literature, sometimes under the name P-values (Meng (1994) and Bayarri and Berger (2000)) and on other occasions under the name *measures of surprise* (Bayarri and Berger (1999)). Usefulness in many contexts is not suspect. In this paper, from the calibration point of view, we propose a new decision rule based on the sampling distribution of the Bayes factor. This rule has some interesting properties, and hence can be considered a serious alternative to other traditional default rules which are based purely on the observed Bayes factor. More detailed discussion regarding this issue is in Section 4.2.

The use of the sampling distribution has been considered in other situations. Box (1980) considered the prior predictive probability of the observed marginal likelihood as a measure for "an overall predictive check" under the model being entertained. Ibrahim, Chen and Sinha (2001) used this approach to calibrate a Bayesian criterion, called the L measure, for model assessment and model comparison. There the prior predictive distribution of the L measure statistic under the true model is defined to be the *calibration distribution.* We borrow this term to denote the prior predictive distribution of the Bayes factor under each of the two competing models.

In this paper, the properties of the calibration distribution of the Bayes factor under each model are studied in detail. We observe that the calibration distribution is far from being symmetric across models. Moreover, this asymmetry depends on the models being compared. As a consequence, a decision rule for determining the strength of evidence based only on the observed value of the Bayes factor is problematic. This problem is also mentioned in Vlachos and

Gelfand (2003). In addition, it is well known that the Bayes factor depends on the choice of the prior distribution and, in particular, it is extremely sensitive to imprecise priors.

We propose a novel calibrating value of the Bayes factor based on the prior predictive distributions and then develop the decision rule, based on this calibrating value, for selecting the model. The calibrating value is motivated by the principle of prior equity and it is computed based on the calibration distributions of the Bayes factor under both models. As we will show, the calibration distribution of the Bayes factor is biased toward the model under which the Bayes factor is being calibrated. We also show that the proposed decision rule based on the calibration distributions is, under mild conditions, equivalent to the surprise-based decision. That is, we choose the model for which the observed Bayes factor is less surprising. Moreover we demonstrate that, in certain situations, the decision rule based on the calibrating value is closely related to the classical rejection region for a standard hypothesis testing problem, and the equivalent classical rejection region is quite robust to the choice of imprecise priors.

The rest of the paper is organized as follows. Section 2 deals with motivation and introduces the notation used throughout. The main results are presented in Section 3. In the same section, a new calibrating value is proposed, and the theoretical properties of calibration distributions for the Bayes factor are examined in detail. Also in Section 3, a new decision rule based on the calibrating value is proposed. In Section 4, we explore the properties of the proposed calibrating value and establish the relationship between the new decision rule and the classical rejection region. In addition, we discuss the relationship between the proposed calibration and the prior model probabilities and we develop a Monte Carlo-based algorithm for computing the calibrating value. In Section 5, we further examine the robustness of the decision rule based on the calibration distribution under both nested and non-nested models. An example using the radiate pine data is given in Section 6 and several related issues and important extensions of the proposed methodology are discussed in Section 7.

2. Motivation and Notation

Assume we are interested in comparing two models, M_1 and M_2, say, as convenient statistical representation of some data $\boldsymbol{y}$. Let $p_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$ and $\pi_i(\boldsymbol{\theta}_i)$ denote the probability function and the (proper) prior distribution, respectively, under model M_i, for $i = 1, 2$. Also, let $m_i(\boldsymbol{y})$ denote the prior predictive distribution under model M_i, that is,

$$m_i(\boldsymbol{y}) = \int p_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\, \pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \qquad i = 1, 2.$$

In the context of hypothesis testing, the selection between these two models can be expressed as

$$H_1 : M_1 \text{ is true}, \quad \text{vs.} \quad H_2 : M_2 \text{ is true}. \tag{1}$$

Given the observed data, $\boldsymbol{y}$, the Bayes factor for M_1 against M_2 is $B_{12}(\boldsymbol{y}) = m_1(\boldsymbol{y})/m_2(\boldsymbol{y})$. It is often not available in closed form. Therefore, numerical or Monte Carlo approximations are needed. Due to recent computational advances, sophisticated techniques for computing Bayes factors have been developed. See, for example, Kass and Vaidyanathan (1992), Chib (1995), Meng and Wong (1996), Chen and Shao (1997), DiCiccio, Kass, Raftery and Wasserman (1997), Gelman and Meng (1998) and Chib and Jeliazkov (2001).

Note that, if $B_{12}(\boldsymbol{y}) > 1$, then $\boldsymbol{y}$ is best predicted by M_1 and consequently, this model is preferred. Of course, using similar arguments, $B_{12}(\boldsymbol{y}) < 1$ gives support to model M_2. Several different ways have been proposed to interpret the strength of evidence according to $B_{12}(\boldsymbol{y})$. Jeffreys (1961) proposed the rule given in Table 1 and, more recently, Kass and Raftery (1995) proposed a slight modification of Jeffreys' proposal as shown in Table 2. Although these rules are given in terms of *evidence against* M_2, the same rule (at least in principle) can be used to interpret the evidence against M_1 by inverting the value of the Bayes factor.

Table 1. Jeffreys' scale of evidence.

$\mathrm{B}_{12} \in$	Evidence against M_2
(1,3.2)	Not worth more than a bare mention
(3.2,10)	Substantial
(10,100)	Strong
(100,∞)	Decisive.

Table 2. Scale of evidence proposed by Kass and Raftery (1995).

$\mathrm{B}_{12} \in$	Evidence against M_2
(1,3)	Not worth more than a bare mention
(3,20)	Positive
(20,150)	Strong
(150,∞)	Very strong.

Prior to the sample, the Bayes factor is a function of the random vector $\boldsymbol{Y}$. The sample distribution of $B_{12}(\boldsymbol{Y})$ under model M_i is defined by $m_i(\cdot)$, the marginal prior distribution of $\boldsymbol{Y}$ under M_i. To investigate whether the above mentioned scale of evidence can be interpreted under the calibration distribution, we use an elementary example below, also considered by Vlachos and Gelfand (2003).

Example 1. Suppose $y_1, \ldots, y_n$ is a sample from model M_i $(i = 1, 2)$, where M_1 is $N(\theta_0, \sigma^2)$ with θ_0 and σ^2 known, and M_2 is $N(\theta, \sigma^2)$ with $\theta \sim N(0, \tau^2)$, and again assume σ^2 and τ^2 are known. The sampling distribution of the Bayes factor is related to the chi-square distribution (under both models), and the characteristics, such as means and tail probabilities, of this distribution can be easily calculated. As an illustration, let $n = 20$, $\theta_0 = 0$, $\sigma^2 = 1$ and $\tau^2 = 5$. After some simple algebra, the Bayes factor can be expressed as

$$B_{12} = \exp\left\{\frac{1}{2}\left(\ln(1+a) - \frac{na}{\sigma^2(1+a)}\bar{y}^2\right)\right\},$$

where $a = n\tau^2/\sigma^2$. It is easy to observe that the range of B_{12} is $(0, (1+a)^{1/2})$. Figure 1 shows the density of $B_{12}(\boldsymbol{Y})$ under M_1 (left) and under M_2 (right).

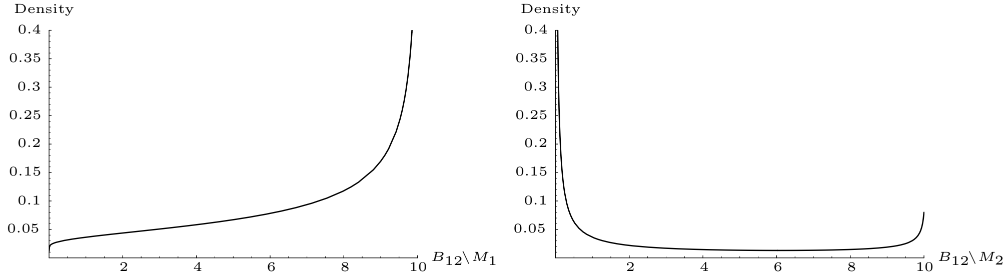

Figure 1. Densities of $B_{12}(\boldsymbol{Y})$ under M_1 (left) and under M_2 (right).

Based on the scale of evidence in Table 1, the probability of (at least) strong evidence against M_2 if M_1 is the true model is 0.08. But, the probability of (at least) strong evidence against M_1 if M_2 is the true model is 0.76. Thus, if the sampling distribution of $\boldsymbol{Y}$ is considered, the original $B_{12}(\boldsymbol{y})$ for fixed data is no longer proper and calibration is needed.

Next, we discuss what would happen if we want to select only one model. Suppose that, as usual, we decide to select M_1 if $B_{12}(\boldsymbol{y}) > 1$ and select M_2 otherwise. The probability of selecting M_2 given that M_1 is true is 0.03, and the probability of selecting M_1 given that M_2 is true is 0.17. According to this, it is more likely to make a mistake in selecting M_1 than M_2. Note that these two probabilities are closely related to the type II error (or to the power function) probabilities in a frequentist sense. Nevertheless, in the Bayesian framework it is possible to evaluate *both* errors, which is often more difficult from the frequentist point of view.

According to the Example 1, and in judicial terms, the Bayes factor is a judge predisposed in favor one of the models *if* the scale of evidence given either in Table 1 or Table 2 is used. Since the calibration distribution of the Bayes factor is asymmetric, a symmetric scale of evidence should not be used. Here $B_{12} \approx 10$ ($\bar{y} \approx 0$) would *provide* evidence supporting M_1 comparable to evidence supporting M_2 if $B_{12} \approx \gamma$ where $Pr(B_{12} \leq \gamma \mid M_2) = Pr(B_{12} \geq 10 \mid M_1) = 0.08$. Note further that the median of the calibration distribution of $B_{12}(\boldsymbol{Y})$ under M_1 is 8, while the median of $B_{21}(\boldsymbol{Y})$ under M_2 is $e^{20.44}$. A 50% credible interval for $B_{12}(\boldsymbol{Y})$ under M_1 is $(5.22, 9.56)$, while the same interval for $B_{21}(\boldsymbol{Y})$ under M_2 is $(15.94, e^{63.86})$. Moreover, the 90% credible intervals are $(1.50, 10.03)$ and $(0.12, e^{189.77})$, respectively. It is clear that the distribution of $B_{21}(\boldsymbol{Y})$ under M_2 is very flat, which implies that almost every value of B_{12} is equally possible. Thus, at least for this example, there is severe asymmetry in the strength of evidence between these two calibration distributions. This asymmetry casts doubt on the use of a predetermined scale for measuring the strength of evidence solely based on the value of the Bayes factor. Furthermore the use of a critical value, such as 1, of the Bayes factor for determining which model is more favorable is particularly problematic. We propose a new calibrating value of the Bayes factor, as well as new decision rules which generically combine the calibration distributions and the observed value of the Bayes factor, in the next section.

3. Main Results

Surprise measures are usually related to the notion of a P-value. To see how this kind of measure can be related to the Bayes factor, write

$$p_i^L = Pr(B_{12}(\boldsymbol{Y}) \leq B_{12}(\boldsymbol{y}) \mid M_i) \;\text{ and }\; p_i^R = Pr(B_{12}(\boldsymbol{Y}) \geq B_{12}(\boldsymbol{y}) \mid M_i),$$

where $Pr(\cdot \mid M_i)$ stands for the prior predictive distribution of $B_{12}(\boldsymbol{Y})$ under model M_i for $i = 1, 2$. Also define $p_i^* = p_i^L$ if $B_{12}(\boldsymbol{y}) \leq 1$ and $p_i^* = p_i^R$ if $B_{12}(\boldsymbol{y}) > 1$. Finally, let $p_i = \min\{p_i^L, p_i^R\}$. We note that p_i is a two-sided P-value of the Bayes factor B_{12} under model M_i. We also note that Vlachos and Gelfand (2003) considered a one-sided P-value as the surprise measure. In the context of surprise measures, the values p_i are of major interest.

In advance, we assume that the variable $B_{12}(\boldsymbol{Y})$ has a continuous distribution $m_i(\cdot)$ for $i = 1, 2$. Obviously, $p_i^L = 1 - p_i^R$.

Lemma 1. *$B_{12}(\boldsymbol{y}) > 1$ if and only if $p_1^* > p_2^*$.*

Proof. Assume first that $B_{12}(\boldsymbol{y}) > 1$. Note that $p_i^* = \int_{A(B_{12}(\boldsymbol{y}))} m_i(\boldsymbol{t})\, d\boldsymbol{t}$ where $A(B_{12}(\boldsymbol{y})) = \{\boldsymbol{t} \in \mathcal{Y} : B_{12}(\boldsymbol{t}) \geq B_{12}(\boldsymbol{y})\}$ and where $\mathcal{Y}$ denotes the set of all possible values for $\boldsymbol{y}$. Note that if $\boldsymbol{t} \in A(B_{12}(\boldsymbol{y}))$, then

$$\frac{m_1(\boldsymbol{t})}{m_2(\boldsymbol{t})} \geq \frac{m_1(\boldsymbol{y})}{m_2(\boldsymbol{y})} = B_{12}(\boldsymbol{y}) > 1,$$

and for all $\boldsymbol{t} \in A(B_{12}(\boldsymbol{y}))$, $m_1(\boldsymbol{t}) - m_2(\boldsymbol{t}) > 0$, so $p_1^* - p_2^* > 0$.

We proceed by contradiction in the other direction. Assume that $p_1^* > p_2^*$, but $B_{12}(\boldsymbol{y}) \leq 1$. Then, $p_1^* - p_2^* = \int_{\bar{A}(B_{12}(\boldsymbol{y}))} (m_1(\boldsymbol{t}) - m_2(\boldsymbol{t}))\, d\boldsymbol{t}$, where $\bar{A}(B_{12}(\boldsymbol{y})) = \{\boldsymbol{t} \in \mathcal{Y} : B_{12}(\boldsymbol{t}) \leq B_{12}(\boldsymbol{y})\}$. Note that if $\boldsymbol{t} \in \bar{A}(B_{12}(\boldsymbol{y}))$,

$$\frac{m_1(\boldsymbol{t})}{m_2(\boldsymbol{t})} \leq \frac{m_1(\boldsymbol{y})}{m_2(\boldsymbol{y})} = B_{12}(\boldsymbol{y}) \leq 1,$$

and then (similar to above), $p_1^* - p_2^* \leq 0$ which contradicts $p_1^* > p_2^*$.

Unfortunately, Lemma 1 cannot ensure that a large value of the Bayes factor corresponds to a large two-sided P-value p_i, since p_i may not be equal to p_i^* in general. Moreover, it is straightforward to see that $B_{12}(\boldsymbol{y}) > 1$ cannot guarantee $p_1 > p_2$. Nevertheless, Lemma 1 is useful since it implies that the calibration distribution cannot be used to resolve Lindley's paradox. In short, Lindley's paradox says that in Example 1, if we use an imprecise prior for θ under model M_2, the Bayes factor will tend to favor M_1 regardless of $\boldsymbol{y}$. In Example 1, Lindley's paradox says $\lim_{\tau^2 \to \infty} B_{12}(\boldsymbol{y}) = \infty$. In this situation, it can be assumed that $B_{12}(\boldsymbol{y})$ is large enough to ensure that $p_i = p_i^*$ and, using Lemma 1, $p_1 > p_2$. Hence M_1 is preferred over M_2 if an imprecise prior (τ^2 large enough) is used.

Definition 1. If c is nonnegative and $Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) = Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1)$, it is called the calibrating value.

The following theorem shows the existence and uniqueness of the calibrating value c.

Theorem 1. *Assume that* $m_i(\boldsymbol{y}) > 0$ *for all* $\boldsymbol{y}$. *Then, there exists a unique* c *so that*

$$Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) = Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1). \tag{2}$$

Proof. Let $f_{B|i}(b)$ be the density function of $B_{12}(\boldsymbol{Y})$ under model M_i evaluated at b. Also let $F_{B|i}(b)$ denote the cumulative distribution function of $B_{12}(\boldsymbol{Y})$ under model M_i evaluated at b. Now (2) can be expressed as $\int_c^\infty f_{B|2}(b)\, db = 1 - \int_c^\infty f_{B|1}(b)\, db$ or, equivalently, $F_{B|2}(c) + F_{B|1}(c) = 1$. Since $g(\cdot) = F_{B|2}(\cdot) + F_{B|1}(\cdot)$ is continuous and strictly increasing in $[0, \infty)$ with $\lim_{b\to\infty} g(b) = 2$ and $\lim_{b \to 0} g(b) = 0$, the result follows.

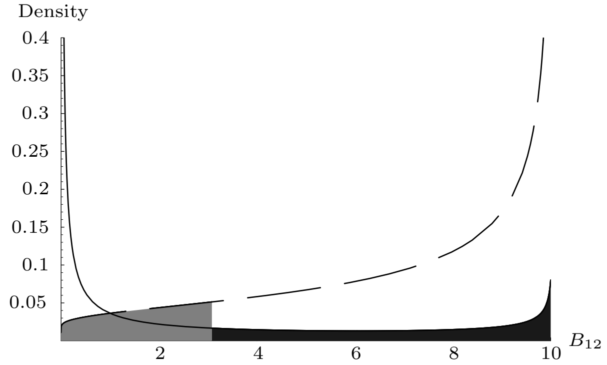

Figure 2. Densities of $B_{12}(\boldsymbol{Y})$ under M_1 (dashed line) and under M_2 (solid line).

As an illustration, the calibrating value is plotted in Figure 2 for Example 1. The value c is 3.08, which makes the black (M_2) and gray (M_1) areas equal. Note that c can be considered as a balanced value so that the error probabilities under models M_1 and M_2 are equal if c is taken to be the *critical value*: we choose M_1 if $B_{12}(\boldsymbol{y})$ is greater than or equal to this value, and we choose M_2 otherwise. As we will show, this value can be used to develop a new decision rule for model selection based on the observed value of the Bayes factor.

Suppose we follow the rule that we select M_1 if $B_{12}(\boldsymbol{y}) > 1$. Then if the $c > 1$, the rule based on $B_{12}(\boldsymbol{y}) > 1$ is biased to M_1, i.e., the error probability under M_1 is less than that under M_2. On the other hand, if $c < 1$, this rule is biased to M_2. In Example 1, $c \approx 3.08$. If we use the c as the critical value, then both error probabilities are the same, namely 0.12. It is interesting to note that this tendency of favoring the simpler model increases with the sample size. If $\sigma = 1, \theta_0 = 0, \tau^2 = 5$ (as before) but $n = 200$, then $c \approx 4.60$ and the error probabilities decrease (as expected) to 0.05.

Theorem 2. *Let μ_1 (assuming it exists) and μ_2 be the means of the calibration distributions of $B_{12}(\boldsymbol{Y})$ under model M_1 and M_2 respectively. Then $\mu_2 = 1 \leq \mu_1$.*

Proof. Note that

$$\mu_i = E(B_{12}(\boldsymbol{Y}) \mid M_i) = E(m_1(\boldsymbol{Y})/m_2(\boldsymbol{Y}) \mid M_i) = \int m_1(\boldsymbol{t}) m_i(\boldsymbol{t})/m_2(\boldsymbol{t}) d\boldsymbol{t}, \ i=1,2.$$

It is clear that $\mu_2 = 1$. To prove the inequality, note that since the function $g(B) = 1/B$ is convex if $B > 0$, Jensen's inequality gives $\mu_1 = E(B_{12}(\boldsymbol{Y}) \mid M_1) \geq [E(1/B_{12}(\boldsymbol{Y}) \mid M_1)]^{-1} = 1$.

Thus the calibration distribution of Bayes factor is biased toward the model under which the Bayes factor is calibrated. The next theorem characterizes the relationship between the medians of the two calibration distributions.

Theorem 3. *Let ξ_i be the median of the calibration distribution of $B_{12}(\boldsymbol{Y})$ under model M_i, for $i = 1, 2$. Then $\xi_2 \leq \xi_1$.*

Proof. Assume first that $\xi_1 \geq 1$, then $Pr(B_{12}(\boldsymbol{Y}) \geq \xi_1 \mid M_1) = 0.5$, or equivalently $\int_{A(\xi_1)} m_1(\boldsymbol{t})\, d\boldsymbol{t} = 0.5$, where $A(\xi_1) = \{\boldsymbol{t} \in \mathcal{Y} : B_{12}(\boldsymbol{t}) \geq \xi_1\}$. Note that if $\boldsymbol{t} \in A(\xi_1)$ then $m_1(\boldsymbol{t}) \geq \xi_1\, m_2(\boldsymbol{t}) \geq m_2(\boldsymbol{t})$, so $0.5 = \int_{A(\xi_1)} m_1(\boldsymbol{t})\, d\boldsymbol{t} \geq \int_{A(\xi_1)} m_2(\boldsymbol{t})\, d\boldsymbol{t}$, and, in consequence, $Pr(B_{12}(\boldsymbol{Y}) \geq \xi_1 \mid M_2) \leq 0.5$ and $\xi_2 \leq \xi_1$. The proof is similar when $\xi_1 < 1$.

A direct consequence of Theorem 3 and the definition of c is that $\xi_2 \leq c \leq \xi_1$. Hence the probabilities in (2) cannot exceed 0.50.

Unlike the means of the calibration distributions, it may not be always true that $\xi_2 \leq 1 \leq \xi_1$. However, this condition is crucial in order to establish the relationship between P-values and the observed value of the Bayes factor.

Definition 2. We say that models M_1 and M_2 are enough separated a priori, if $\xi_2 \leq 1 \leq \xi_1$.

To get an idea about how much the two models can be close to each other so that the condition given in Definition 2 is not satisfied, we revisit Example 1.

Example 1.(continued) For ease of exposition, we assume that $\theta_0 = 0$. Then for τ^2 small M_2 approximates M_1. How small can τ^2 be so that the condition in Definition 2 does not hold? It can be shown that τ^2 must be less than 0.15 in order that the condition of *enough separation a priori* does not hold.

Theorem 4. *For selecting between models M_1 and M_2, let c denote the calibrating value and suppose that these two models are enough separated a priori. Then $B_{12}(\boldsymbol{y}) > c$ if and only if $p_1 > p_2$.*

Proof. Suppose first that $B_{12}(\boldsymbol{y}) > c$. Since Theorem 3 holds, there are only two cases to consider:

Case 1. $B_{12}(\boldsymbol{y}) > \xi_1$. In this case, it is easy to show that $p_i = p_i^*$, $i = 1, 2$. Moreover, since $B_{12}(\boldsymbol{y}) > \xi_1 \geq 1$, then following Lemma 1, $p_1^* > p_2^*$, so $p_1 > p_2$.

Case 2. $c < B_{12}(\boldsymbol{y}) \leq \xi_1$. In this case, it is straightforward to see that $p_1 = p_1^L$ and $p_2 = p_2^R$, while $p_2^R < Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) = Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1) < p_1^L$, so $p_1 > p_2$.

We prove the other direction by contradiction. Suppose that $p_1 > p_2$, but $B_{12}(\boldsymbol{y}) \leq c$. Since Theorem 3 holds, there are only two cases to consider.

Case 3. $B_{12}(\boldsymbol{y}) \leq \xi_2$. In this case $p_i = p_i^*$, $i = 1, 2$. Moreover, since $B_{12}(\boldsymbol{y}) \leq \xi_2$ ≥ 1, then following Lemma 1, $p_1^* \leq p_2^*$, so $p_1 \leq p_2$, which contradicts the hypothesis.

Case 4. $\xi_2 < B_{12}(\boldsymbol{y}) \leq c$. In this case, it is straightforward to see that $p_1 = p_1^L$ and $p_2 = p_2^R$, but $p_2^R \leq Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) = Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1) \geq p_1^L$, so $p_1 \geq p_2$, which contradicts the assumption.

Suppose that a large value of $B_{12}(\boldsymbol{y})$ implies that M_1 is better in fitting the data $\boldsymbol{y}$ than M_2. We are led to the following decision rule.

Rule 1. Select model M_1 as true if $B_{12}(\boldsymbol{y}) > c$ and select model M_2 otherwise.

On the other hand, from the surprise measure point of view, the less the surprise under a model, the more the evidence in favor of the model. This principle leads to the following decision rule.

Rule 2. Select model M_1 if $p_1 > p_2$.

According to Theorem 4 and under its conditions, Rule 1 is equivalent to Rule 2. Interestingly, Rule 1 has some other connections with frequentist procedures, more exactly, with conditional frequentist procedures, see Berger, Brown and Wolpert (1994) and Berger, Boukai and Wang (1997) for details.

4. Properties and Computation of Calibrating Value

4.1. Calibrating value and rejection region

To decide whether a (null) hypothesis (or a model) should be rejected, in the frequentist framework, a rejection region (FRR) is defined, so that if $\boldsymbol{y}$ lies in FRR, the null hypothesis is rejected.

From the calibrating point of view, the rule "select M_2 as true if $B_{12}(\boldsymbol{y}) < c$", can be rewritten as "reject H_1 if $B_{12}(\boldsymbol{y}) < c$". So $B_{12}(\boldsymbol{y}) < c$ can be also viewed as a Rejection Region, denoted by CRR, but differences remain. First, the CRR is constructed under both models. Second, to construct FRR, the probability of type I error is usually prespecified. It is usually felt that it should be a function of the sample size (see Berger and Sellke (1997) and Bayarri and Berger (1999)). In any case, its imposition is problematic and certain standard levels of significance are overused. To construct CRR, none of this is needed since the calibrating value c is computed from the prior equity of the error probabilities under both models. Interestingly, the probability of type I error will then depend on the sample size.

Example 1.(continued) For simplicity we consider the case in which $\theta_0 = 0$. In terms of the statistic $\bar{y}$, it is easy to show that, if $c = c(\sigma, \tau^2, n)$ is the calibrating value, CRR has the following form

$$CRR = \left\{ \boldsymbol{y} \in \mathcal{Y} \,:\, |\bar{y}| > \frac{\tau}{a}\sqrt{(1+a)\ln\frac{1+a}{c^2}} \right\},$$

where $a = n\tau^2/\sigma^2$. The FRR for testing H_1: $\theta = 0$ against H_2: $\theta \neq 0$ is

$$FRR = \left\{ \boldsymbol{y} \in \mathcal{Y} \,:\, |\bar{y}| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\},$$

where $0 < \alpha < 1$ is a prespecified level of significance. Figure 3 shows c and $r = (\tau/a)\sqrt{(1+a)\ln\left[(1+a)/c^2\right]}$ as a function of τ^2 for fixed $\sigma = 1$ and $n = 20$, for $\tau^2 \in (0.15, 10,000)$. From Figure 3 (left), we can see that c increases with τ^2. As discussed earlier, when τ^2 grows, the prior becomes more imprecise and, as a result, the Bayes factor tends to be biased toward model M_1. Also, for a fixed $\bar{y}$, the Bayes factor increases with τ^2. Thus, if we use the scale of evidence listed either in Table 1 or Table 2, model M_1 is highly favored for large τ^2. However, based on our proposed Rule 1, we choose M_1 only if $B_{12}(\boldsymbol{y}) > c$. For instance, when $\tau^2 = 1,000$, $c \approx 6$. Then we select M_1 only if $B_{12}(\boldsymbol{y}) > 6$. From Figure 3 (right), the r curve is relatively flat. This implies that the CRR is relatively robust to the choice of τ^2. This result demonstrates that our proposed decision rule is more robust to the choice of an imprecise prior than those using the scale of evidence constructed based on the value of the Bayes factor. Finally we mention that, for a given α, CRR matches FRR for a certain τ^2. This is interesting, since if we use this equality to construct a prior distribution, then Bayesian and frequentist answers agree.

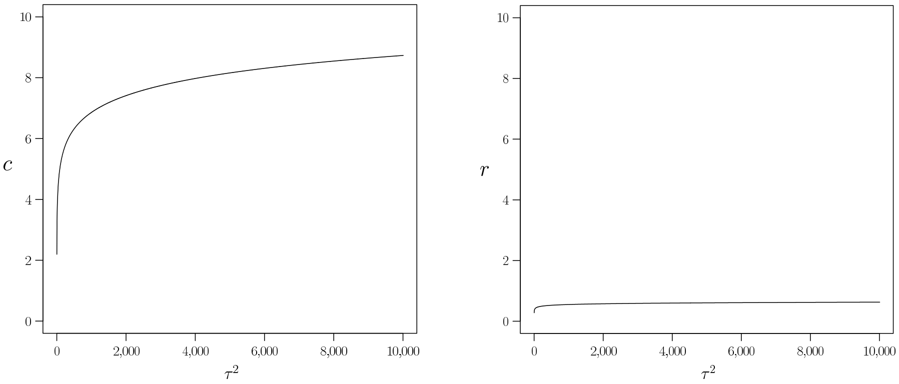

Figure 3. Behavior of c (left) and r_c (right) as a function of τ^2 for Example 1 ($\sigma = 1, n = 20, \theta_0 = 0$).

4.2. Calibration and prior model probabilities

When two models are compared, the *Bayes action* corresponding to a 0-l_i function loss (see Berger (1985) for proper definitions) is to select M_1 if

$$B_{12}(\boldsymbol{y}) > \frac{\psi_2}{\psi_1}\frac{l_1}{l_2},$$

where ψ_i is the prior probability of M_i being the true model and l_i is the loss associated with the wrong decision, $i = 1, 2$.

It would be very difficult to argue that losses can be determined by intrinsic characteristics of the models under comparison since the losses l_i depend on the specific nature of the problem at hand. Hence, we assume that l_i are known. Without loss of generality, we take $l_1 = l_2$. It is straightforward to see that under this decision rule, we select the model that has the larger posterior probability. Then, using c as the critical value is equivalent to assigning c to the quotient ψ_2/ψ_1, so $\psi_1 = 1/(1+c)$ and $\psi_2 = c/(1+c)$. We do not think that this assignment should be used to substitute for an actual elicitation of the ψ_i. Instead, we consider this assignment as a default elicitation of prior model probability when the ψ_i's are unknown.

Little attention has been paid to default assignments of prior model probabilities compared to the considerable effort that has been given to assigning default prior distributions (see Berger and Pericchi (2001) for an excellent overview of default prior distributions).

Next, we compare the default assignment (ψ_i^c) based on the calibrating value with two default assignments of prior probabilities: equal prior model probability assignment (ψ_i^e) and the assignment (ψ_i^s) from Bayesian hypothesis testing. Since we only deal with two competing models, we have $\psi_i^e = 0.5$, $i = 1, 2$. Given a model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ and a proper prior distribution, $\pi(\boldsymbol{\theta})$, the two competing models are $p_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i) = \{p(\boldsymbol{y} \mid \boldsymbol{\theta}_i),\ \ \boldsymbol{\theta}_i \in \Theta_i\}$, $i = 1, 2$. Then, $\psi_i^s = \int_{\Theta_i} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ for $i = 1, 2$. We note that ψ_i^e and ψ_i^c are always well defined while ψ_i^s is defined only in a few cases. Although the rule based on ψ_i^e is attractive, it implies the comparison of Bayes factor with fixed critical values. Apart from the frequentist consequences as discussed in Section 2, Lavine and Schervish (1999) showed that fixed critical values can lead to an incoherent procedure.

Example 2. Let $p(y \mid \theta)$ be $N(\theta, 1)$, and consider the following models M_1: $p_1(y \mid \theta_1) = \{p(y \mid \theta_1),\ \ \theta_1 \in \Theta_1\}$ and M_2: $p_2(y \mid \theta_2) = \{p(y \mid \theta_2),\ \theta_2 \in \Theta_2\}$. The prior distributions are $\pi_1(\theta_1) \propto N(\theta_1 \mid \mu, 1) 1_{\Theta_1}(\theta_1)$ and $\pi_2(\theta_2) \propto N(\theta_2 \mid \mu, 1) 1_{\Theta_2}(\theta_2)$, which come from the *common* prior distribution $\pi(\theta \mid \mu) = N(\theta \mid \mu, 1)$.

If $\Theta_1 = [0, \infty)$ and $\Theta_2 = (-\infty, 0)$, it is straightforward to see that Bayes factor (expressed in terms of μ) is

$$B_{12}(\mu) = \left(\frac{1 - \Phi(-\dfrac{y+\mu}{\sqrt{2}})}{\Phi(-\dfrac{y+\mu}{\sqrt{2}})} \right) \left(\frac{1 - \Phi(\mu)}{\Phi(\mu)} \right),$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

If $\mu = 0$, the assignment ψ_i^e seems to be justified, since the models are clearly balanced. Interestingly, in this case, $\psi_i^c = \psi_i^s = 0.5$, also. If for instance $\mu = 0.5$, then $\psi_1^c = 0.48$ and $\psi_1^s = 0.69$. If $\mu = 2$, $\psi_1^c = 0.44$ and $\psi_1^s = 0.98$. It seems that assignment ψ_i^c is closer to 0.5 than ψ_i^s, which takes large values quickly. This fact is also shown more precisely in Figure 4 (left).

Example 3. For the two competing models of Example 2, consider now the subspaces $\Theta_1 = (-k, k)$ and $\Theta_2 = (-\infty, -k] \bigcup [k, \infty)$ for some $k > 0$. The Bayes factor (in terms of k) is

$$B_{12}(k) = \left(\frac{2-2\Phi(k)}{2\Phi(k)-1}\right)\left(\frac{\Phi((k-\frac{y}{2})\sqrt{2}) - \Phi(-(k+\frac{y}{2})\sqrt{2})}{1+\Phi(-(k+\frac{y}{2})\sqrt{2}) - \Phi((k-\frac{y}{2})\sqrt{2})}\right).$$

If $k = 0.6745$, then $\psi_1^s = 0.5$ and $\psi_1^c = 0.559$. If $k = 1$, then $\psi_1^s = 0.683$ and $\psi_1^c = 0.547$ and if $k = 0.1$, $\psi_1^s = 0.080$, $\psi_1^c = 0.550$. Figure 4 (right) shows the behavior of ψ_i as a function of $k \in [0.10, 1.25]$ for the different assignments.

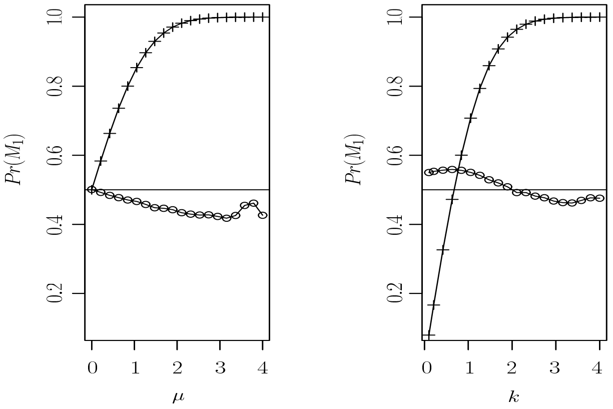

Figure 4. $Pr(M_1)$ as a function of μ (left) and k (right). The line with "o" corresponds to ψ_1^c; that with "+" corresponds to ψ_1^s; the line without symbol corresponds to ψ_1^e

From Examples 2 and 3, we can see that ψ_i^c is fairly close to ψ_i^e, while ψ_i^s can easily become very large. Based on our experience with calibrating values, when the dimensions of the parameters in the two models are similar, 0.50 is a valid approximation to ψ_i^c; when the two dimensions are substantially different, ψ_i^c can be very different from 0.50.

4.3. Computation

Let

$$S(u) = \int_{\bar{A}(u)} (m_1(\boldsymbol{t}) + m_2(\boldsymbol{t}))d\boldsymbol{t}, \tag{3}$$

where $\bar{A}(u) = \{\boldsymbol{t} \in \mathcal{Y} : B_{12}(\boldsymbol{t}) \leq u\}$. Then the calibrating value c is the unique value such that $S(c) = 1$. In general, the closed form expression for $S(u)$ is not available. Thus, we will develop a Monte Carlo (MC) procedure to find the solution of this integral equation.

Let $B_{12}^{j|i}$ denote independent or ergodic simulated observations of $B_{12}(\boldsymbol{Y})$, with $\boldsymbol{Y}$ having distribution $m_i(\boldsymbol{y})$ for $i = 1, 2$ and $j = 1, \ldots, N$. Also, let $B_{12}^{(1)} \leq$

$\ldots \leq B_{12}^{(2N)}$ denote the ordered values of the $B_{12}^{j|i}$. Then an MC approximation to function $S(\cdot)$ is $\tilde{S}(\cdot)$, where

$$\tilde{S}(u) = \begin{cases} 0 & \text{if } \ u < B_{12}^{(1)}, \\ r/N & \text{if } \ B_{12}^{(r)} \leq u < B_{12}^{(r+1)}, \\ 2 & \text{if } \ u \geq B_{12}^{(2N)}. \end{cases}$$

Thus, $\tilde{S}(c) = 1$ if and only if $B_{12}^{(N)} \leq u < B_{12}^{(N+1)}$, so all c values in $[B_{12}^{(N)}, B_{12}^{(N+1)})$ are approximate MC solutions to (3). In order to have a unique value, we suggest taking

$$c = (B_{12}^{(N)} + B_{12}^{(N+1)})/2 \tag{4}$$

as the MC solution. Note that this solution is subject to a maximum error $\delta = (B_{12}^{(N+1)} - B_{12}^{(N)})/2$. To reduce computational effort, a bound on the maximum error may be determined. An initial solution of c from an initial sample of size N can be computed, N augmented if the desired error bound is not attained.

It will be usually the case that the analytic form of the distribution of $B_{12}(\boldsymbol{Y})$ is not available (under one or both models). Thus, sampling $B_{12}^{j|i}$ directly from the distribution of $B_{12}(\boldsymbol{Y})$ under model M_i is an impossible task. However, the following algorithm can be used.

Step 1. Generate $\boldsymbol{\theta}_i^j$ from $\pi_i(\boldsymbol{\theta}_i)$.
Step 2. Given $\boldsymbol{\theta}_i^j$, generate $\boldsymbol{t}_i^j$ from $p(\cdot \mid \boldsymbol{\theta}_i^j, M_i)$.
Step 3. Compute $B_{12}^{j|i} = m_1(\boldsymbol{t}_i^j)/m_2(\boldsymbol{t}_i^j)$.

We repeat the above algorithm for $i = 1, 2$ and $j = 1, \ldots, N$.

5. Robustifying Bayes Factor Based Decisions

It has been argued that Bayes factor is very sensitive to the prior information, especially in a context of vague prior information (see for instance O'Hagan (1994, p.193)). This affirmation is based on the fact that Bayes factors are affected by even little changes in the prior distribution representing vague information.

Usually, vague prior information is modeled via proper distributions governed by a parameter, say τ, in which, large values of τ represent vague information. This parameter is closely related to the prior variance. This may be unwise. Consider the following three examples.

Example 4. Suppose $M_1 : Y \sim \text{Exp}(1)$ and $M_2 : Y \sim \text{Exp}(\theta)$ ($\text{Exp}(\theta)$ stands for exponential distribution with mean θ^{-1}) and suppose that, roughly, θ has a priori, mean μ. In order to assign a prior distribution (under M_2) in such a context, we take a Gamma distribution with mean μ and *large* variance τ, that is, $\pi_2(\theta) = \mathrm{Ga}(\theta \mid \alpha = \mu^2/\tau,\ \beta = \mu/\tau)$.

Straightforward algebra shows that the marginal distributions are, $m_1(y) = e^{-y}$ and

$$m_2(y) = \mu \left(\frac{y\tau}{\mu} + 1 \right)^{-(\frac{\mu^2}{\tau}+1)}.$$

The Bayes factor, as a function of τ, is given by

$$B_{12}(\tau) = \frac{e^{-y}}{\mu} \left(\frac{y\tau}{\mu} + 1 \right)^{\frac{\mu^2}{\tau}+1}.$$

If another magnitude of vagueness, say $\tau\nu$ with $\nu > 0$, is chosen, then

$$B_{12}(\tau\nu) = \frac{e^{-y}}{\mu} \left(\frac{y\tau\nu}{\mu} + 1 \right)^{\frac{\mu^2}{\tau\nu}+1}.$$

The quotient between these two quantities is

$$\frac{B_{12}(\tau\nu)}{B_{12}(\tau)} = \nu\, f(\tau), \qquad f(\tau) = \frac{\left(\frac{y\tau}{\mu} + \frac{1}{\nu}\right) \left(\frac{y\tau\nu}{\mu} + 1\right)^{\frac{\mu^2}{\tau\nu}}}{\left(\frac{y\tau}{\mu} + 1\right) \left(\frac{y\tau}{\mu} + 1\right)^{\frac{\mu^2}{\tau}}}.$$

It is easy to show that $f(\tau) = 1 + o(1)$ and hence $B_{12}(\tau\nu)/B_{12}(\tau) = \nu(1 + o(1))$. When τ is large, representing vague information, $B_{12}(\tau\nu) \approx \nu B_{12}(\tau)$. Consequently, the Bayes factor is very sensitive to the choice of the value representing vagueness.

Example 5. Let $p(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid X_1\boldsymbol{\beta}_1 + X_e\boldsymbol{\beta}_e, \sigma^2 I_n)$, where X_1 and X_e are $n \times k_1$ and $n \times k_e$ matrices, and $X_2 = [X_1 : X_e]$ is an $n \times k_2$ matrix with $k_2 = k_1 + k_e$. For simplicity, we assume that X_2 is of full rank. $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_e$ are k_1 and k_e-dimensional respectively. Take M_1: $p_1(\boldsymbol{y} \mid \boldsymbol{\phi}_1, \sigma) = p(\boldsymbol{y} \mid \boldsymbol{\beta}_1 = \boldsymbol{\phi}_1, \boldsymbol{\beta}_e = \mathbf{0}, \sigma)$ and M_2: $p_2(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = p(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma)$. This is a typical *variable selection* problem.

The prior distributions are $\pi_1(\boldsymbol{\phi}_1, \sigma_1)$ and $\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma_2) = \pi_{2.1}(\boldsymbol{\beta}_1, \sigma_2)\ \pi_{2.2}(\boldsymbol{\beta}_e)$, where π_1, $\pi_{2.1}$ and $\pi_{2.2}$ are proper distributions. Let us focus on $\pi_{2.2}$ and suppose that vague prior information allows assigning $\pi_{2.2}(\boldsymbol{\beta}_e) = N_{k_e}(\boldsymbol{\beta}_e \mid \boldsymbol{\beta}_e^0, \tau\Sigma)$, but with τ large (modeling vagueness). It is straightforward to show that, as τ becomes large,

$$\pi_{2.2}(\boldsymbol{\beta}_e) = (\frac{1}{2\pi\tau})^{\frac{k_e}{2}} (\det(\Sigma))^{-1/2} \cdot (1 + o(\tau^{-l})), \tag{5}$$

with $l < 1$.

The marginal distribution under M_2 is

$$m_2(\boldsymbol{y}) = \int \left(\frac{1}{2\sigma_2^2\pi}\right)^{\frac{n-k_e}{2}} \exp\{-\frac{1}{2\sigma_2^2}(\boldsymbol{y} - X_1\boldsymbol{\beta}_1)'(I - P_e)(\boldsymbol{y} - X_1\boldsymbol{\beta}_1)\}$$
$$\times\det(X_e'X_e)^{1/2}\left(\frac{1}{2\sigma_2^2\pi}\right)^{\frac{k_e}{2}} \exp\{-\frac{1}{2\sigma_2^2}(\boldsymbol{\beta}_e - \tilde{\boldsymbol{\mu}})'(X_e'X_e)(\boldsymbol{\beta}_e - \tilde{\boldsymbol{\mu}})\}$$
$$\times\pi_{2.2}^D(\boldsymbol{\beta}_e)\det(X_e'X_e)^{-1/2}\pi_{2.1}(\boldsymbol{\beta}_1, \sigma_2)d\boldsymbol{\beta}_e d\boldsymbol{\beta}_1 d\sigma_2,$$

with $\tilde{\boldsymbol{\mu}} = (X_e'X_e)^{-1}X_e'(\boldsymbol{y} - X_e\boldsymbol{\beta}_e)$ and $P_e = X_e(X_e'X_e)^{-1}X_e'$.

Plugging in (5) and integrating out with respect to $\boldsymbol{\beta}_e$, we obtain

$$m_2(\boldsymbol{y}) = \left(\frac{1}{2\tau\pi}\right)^{k_e/2}(\det(\Sigma))^{-1/2}m_{2.1}(\boldsymbol{y})(1 + o(\tau^{-l})),$$

with

$$m_{2.1}(\boldsymbol{y}) = \int \left(\frac{1}{2\sigma_2^2\pi}\right)^{\frac{n-k_e}{2}} \exp\{-\frac{1}{2\sigma_2^2}(\boldsymbol{y} - X_1\boldsymbol{\beta}_1)'(I - P_e)(\boldsymbol{y} - X_1\boldsymbol{\beta}_1)\}$$
$$\times\det(X_e'X_e)^{-1/2}\pi_{2.1}(\boldsymbol{\beta}_1, \sigma_2)d\boldsymbol{\beta}_1 d\sigma_2.$$

After some immediate algebra,

$$B_{12}(\tau) = \tau^{k_e/2}(2\pi)^{k_e/2}(\det(\Sigma))^{1/2}\frac{m_1(\boldsymbol{y})}{m_{2.1}(\boldsymbol{y})}(1 + o(\tau^{-l})).$$

If another magnitude of vagueness, say $\tau\nu$ with $\nu > 0$, was chosen then

$$B_{12}(\tau\nu) = (\nu\tau)^{k_e/2}(2\pi)^{k_e/2}(\det(\Sigma))^{1/2}(\frac{m_1(\boldsymbol{y})}{m_{2.1}(\boldsymbol{y})})(1 + o(\tau^{-l})),$$

and $B_{12}(\tau\nu)/B_{12}(\tau) = \nu^{k_e/2}(1+o(\tau^{-l}))$. When τ is large, we see that $B_{12}(\tau\nu) \approx \nu^{k_e/2}B_{12}(\tau)$. The high dependence of the Bayes factor on the vagueness is clear.

Example 6. Consider two non-nested models: M_1: Y ~LN(θ_1, σ^2) and M_2: Y ~Exp(θ_2), with prior distributions $\pi_1(\theta_1)$ =N$(\theta_1 \mid 0, \tau_1^2)$ and $\pi_2(\theta_2)$ =Ga$(\theta_2 \mid \mu^2/\tau_2, \mu/\tau_2)$. Both, τ_1 and τ_2 are assumed to be large.

It is straightforward to show that

$$m_1(y) = \frac{1}{y\tau_1\sqrt{2\pi}}\exp\left\{-\frac{(\log y)^2}{2\sigma^2}(\frac{\tau_1^2}{\sigma^2} + 1)^{-1}\right\}\left(1 + \frac{\sigma^2}{\tau_1^2}\right)^{-1/2},$$
$$m_2(y) = \frac{\mu^2}{y\tau_2 + \mu}\left(\frac{y\tau_2}{\mu} + 1\right)^{-(\mu^2/\tau_2)}.$$

As τ_1 becomes large, it can be seen that $m_1(y) = (1/(y\tau_1\sqrt{2\pi}))(1+o(1))$. Similarly, as τ_2 becomes large, $m_2(y) = (\mu^2/(y\tau_2+\mu))(1+o(1))$. In this situation, the Bayes factor can be approximated by $B_{12}(\tau_1,\tau_2) \approx (y\tau_2+\mu)/(y\tau_1\mu^2\sqrt{2\pi})$. If another scale of imprecise information in π_2 is used, say $\nu\tau_2$, then $B_{12}(\tau_1,\nu\tau_2) \approx (y\nu\tau_2+\mu)/(y\tau_1\mu^2\sqrt{2\pi})$, and $B_{12}(\tau_1,\nu\tau_2)/B_{12}(\tau_1,\tau_2) \approx (y\nu\tau_2+\mu)/(y\tau_2+\mu) \approx \nu$ since τ_2 is assumed to be large. Clearly $B_{12}(\tau_1,\nu\tau_2) \approx \nu B_{12}(\tau_1,\tau_2)$, and again the Bayes factor is impacted in the context of vague prior information.

It can be shown that similar conclusions obtain if other changes in the scale of imprecise information are made, say, changing τ_1 by $\nu\tau_1$ in π_1 or τ_2 by $\nu_2\tau_2$ and τ_1 by $\nu_1\tau_1$ in both priors.

Examples 4, 5 and 6 clearly demonstrate that the way in which the Bayes factor is affected by vagueness is consistent with the way in which the calibrating value is affected. Hence, the decision based on the calibrating value is unaffected by the vagueness of imprecise prior specification. This idea is designed to robustify decisions in a context of little prior information. However, our proposed rule cannot be used in an *extreme* case (for example, the limiting improper uniform prior for a location parameter) because of the Lindley paradox (although, as shown in this paper, the calibrating value mitigates the impact of such an effect).

6. Radiata Pine Data Example

In this example, we use the dataset of Williams (1959), displayed in Table 3, to further illustrate the proposed methodology. The same dataset is in Carlin and Louis (2000).

Table 3 represents the maximum compressive strength parallel to the grain (y_i), the specimen's density (x_i) and the density adjusted for resin content (z_i) obtained from $n = 42$ specimens of radiata pine. It is desired to compare two models M_1 and M_2, where

$$\begin{aligned} M_1 &: Y_i = \alpha_1 + \beta_1(x_i - \bar{x}) + \epsilon_i^1, \qquad && i = 1,\ldots,n, \\ M_2 &: Y_i = \alpha_2 + \beta_2(z_i - \bar{z}) + \epsilon_i^2, \qquad && i = 1,\ldots,n. \end{aligned}$$

The variables ϵ_i^j are i.i.d. from normal distributions with zero mean and variance σ_j^2, respectively. The proposed prior distributions are $\pi_i(\alpha_i) = N(\alpha_i \mid 3{,}000, 10^6)$, $i = 1,2$, $\pi_i(\beta_i) = N(\beta_i \mid 185, 10^4)$, $i = 1,2$, and $\pi_i(\sigma_j^2) = IG(\sigma_j^2 \mid 3,(2{\cdot}300^2))$, $i = 1,2$, where $IG(3,(2{\cdot}300^2))$ stands for an inverse gamma distribution having both mean and standard deviation equal to 300^2. Also assume that

α_i, β_i and σ_i^2 are independent a priori, that is, $\pi_i(\alpha_i, \beta_i, \sigma_i^2) = \pi_i(\alpha_i)\pi_i(\beta_i)\pi_i(\sigma_i^2)$. These proposed prior distributions are very vague although still proper.

Table 3. Measures on specimens of radiata pine. y_i represents the maximum compressive strength parallel to the grain; x_i is the specimen's density and z_i the density adjusted for resin content obtained from $n = 42$ specimens of radiata pine.

Case (i)	y_i	x_i	z_i
1	3040	29.2	25.4
2	2470	24.7	22.2
3	3610	32.3	32.2
4	3480	31.3	31.0
5	3810	31.5	30.9
6	2330	24.5	23.9
7	1800	19.9	19.2
8	3110	27.3	27.2
9	3160	27.1	26.3
10	2310	24.0	23.9
11	4360	33.8	33.2
12	1880	21.5	21.0
13	3670	32.2	29.0
14	1740	22.5	22.0
15	2250	27.5	23.8
16	2650	25.6	25.3
17	4970	34.5	34.2
18	2620	26.2	25.7
19	2900	26.7	26.4
20	1670	21.1	20.0
21	2540	24.1	23.9
22	3840	30.7	30.7
23	3800	32.7	32.6
24	4600	32.6	32.5
25	1900	22.1	20.8
26	2530	25.3	23.1
27	2920	30.8	29.8
28	4990	38.9	38.1
29	1670	22.1	21.3
30	3310	29.2	28.5
31	3450	30.1	29.2
32	3600	31.4	31.4
33	2850	26.7	25.9
34	1590	22.1	21.4
35	3770	30.3	29.8
36	3850	32.0	30.6
37	2480	23.2	22.6
38	3570	30.3	30.3
39	2620	29.9	23.8
40	1890	20.8	18.4
41	3030	33.2	29.4
42	3030	28.2	28.2

To calculate the calibrating value, c, we use the algorithm given in Section 4.3. The Bayes factor is computed using the method of Chib (1995). Let $\boldsymbol{t}$ be a simulated sample from $m_i(\cdot)$, let $\{\alpha_i^{(g)}, \beta_i^{(g)}, (\sigma_i^2)^{(g)}\}_{g=1}^G$ be the Gibbs output obtained from the posterior distribution $\pi_i(\alpha_i, \beta_i, \sigma_i^2 \mid \boldsymbol{t})$ and let $p(\alpha_i, \beta_i \mid \boldsymbol{t}, \sigma_i^2)$ and $p(\sigma_i^2 \mid \boldsymbol{t}, \alpha, \beta)$ be the full posterior conditionals of (α_i, β_i) and σ_i^2 respectively, under M_i, for $i = 1, 2$. Following Chib (1995), the logarithm of the marginal distribution under M_i is approximated by

$$\begin{aligned}\log \hat{m}_i(\boldsymbol{t}) = {} & \log p(\boldsymbol{t} \mid \hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}_i^2, M_i) + \log \pi_i(\hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}_i^2) \\ & - \log p(\hat{\alpha}_i, \hat{\beta}_i \mid \boldsymbol{t}, \hat{\sigma}_i^2) - \log \hat{\pi}(\hat{\sigma}_i^2 \mid \boldsymbol{t}), \qquad i = 1, 2, \end{aligned} \tag{6}$$

where $\hat{\pi}(\hat{\sigma}_i^2 \mid \boldsymbol{t}) = G^{-1} \sum_{g=1}^G p(\hat{\sigma}_i^2 \mid \boldsymbol{t}, \alpha^{(g)}, \beta^{(g)})$. Although the approximation in (6) holds for any $(\hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}_i^2)$, the maximum likelihood estimator or the posterior mean is recommended. The MC estimate of the Bayes factor $B_{12}(\boldsymbol{t})$ associated with $\boldsymbol{t}$ is then $\hat{B}_{12}(\boldsymbol{t}) = \exp\{\log \hat{m}_1(\boldsymbol{t}) - \log \hat{m}_2(\boldsymbol{t})\}$.

To calculate the calibrating value, the maximum error was fixed at $\delta = 0.002$. This bound was attained with a simulated sample of the Bayes factor of size 1,110, ($N = 555$). The median of B_{12} under M_1 (ξ_1) is 124,557.9 while under M_2 (ξ_2) it is 9.6×10^{-6}. The condition of enough prior separation (see Definition 2) is clearly attained. The calibrating value c is 1.02 ± 0.002. Since $B_{12}(\boldsymbol{y}) = 0.0002$ is the observed Bayes factor, model M_2 should be selected.

It is also of interest to compute the values of the error probabilities under both models. For this dataset, we obtain $Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) = Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1) \approx 0.051$. We notice that when the above probabilities are large, the probability of wrong decision increases and, in this case, we should doubt about the validity of the selection decision.

7. Discussion

Based on the calibration distributions of the Bayes factor, the scale of evidence based on the value of the Bayes factor, such as the one proposed by Jeffreys, may be inappropriate. Instead, we have proposed a new decision rule based on intrinsic characteristics of the calibration distributions of the Bayes factor under both models being compared. This decision rule ensures the same error probabilities *a priori* under both models. More interestingly, the calibrating value from which the new decision rule is constructed is determined *a priori*, i.e., it does not depend on the data.

In Section 3, we assume that the calibration distribution is continuous. This assumption can be relaxed. For a discrete distribution, the calibrating value c can be defined as the solution to $\inf_{c \geq 0} |Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2) - Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1)|$.

The calibrating value c arises by imposing equal prior error probabilities under two competing models. However, in some situations the error under one model could be more important than under the other. If this is the case, the calibrating value can be modified accordingly. For example, for a continuous calibration distribution, take $Pr(B_{12}(\boldsymbol{Y}) \leq c \mid M_1) = k_0 Pr(B_{12}(\boldsymbol{Y}) \geq c \mid M_2)$, where $0 < k_0 < 1$, if the error under M_1 is believed to be more serious than under M_2 *a priori*. In this case, our Rule 1 remains the same, but Rule 2 needs to be modified.

We have considered for selecting one of the two competing models. However, calibration distributions can also be used to develop a set of scales for interpreting the strength of evidence provided by the Bayes factor. One possibility is to use the surprise ratio $r = p_1/p_2$. A large value of r indicates strong evidence in favor of model M_1 and, conversely, a small value of r is an indication of the evidence in favor of model M_2. Then the scale of evidence in favor of one of the two competing models can be developed based on the percentiles of the calibration distributions. Since the calibration distribution is model-dependent, these new scales are quite different than those given in Table 1 or Table 2.

We have focused only on two models but can generalize to more than two models. Suppose there are K models being compared simultaneously. As discussed in Vlachos and Gelfand (2003), obtaining all of pairwise calibration distributions of the Bayes factors becomes computationally prohibitive and, even if we did, interpretation becomes very difficult. For the purpose of "screening", we may consider the following alternative. Let $m_k(\boldsymbol{y})$ denote the marginal likelihood for $k = 1, \ldots, K$. Without loss of generality, we assume $m_1(\boldsymbol{y}) = \max\{m_k(\boldsymbol{y})\}$. Then, we propose to calibrate only $K-1$ Bayes factors $B_{1k} = m_1(\boldsymbol{y})/m_k(\boldsymbol{y})$, $k = 2, \ldots, K$. For these $K-1$ Bayes factors, we compute $K-1$ calibrating values. Then model M_1, along with the models selected based on our proposed Rule 2, will be retained for further analysis. This strategy is computationally feasible and it will produce a reasonably good list of candidate models for further consideration.

Finally, we comment that the methodology developed in this paper can also be extended to some other Bayesian model assessment criteria such as the posterior Bayes factor (Aitkin (1991)), the pseudo-Bayes factor (Geisser and Eddy (1979)), and the L measure (Gelfand and Ghosh (1998) and Ibrahim, Chen and Sinha (2001)). Although the calibration idea can be easily applied to the other criteria, the Bayes factor may be more attractive, since the properties of the calibration distributions shown in Theorems 2 to 4 may not be maintained under the other criteria.

Acknowledgements

The authors wish to thank the Editor, an associate editor and two anonymous referees for helpful comments and suggestions which have led to improvement of this article. This work was done while Dr. García-Donato was visiting the Department of Statistics at the University of Connecticut in Spring, 2002. He gratefully acknowledges the hospitality of its faculty, graduate students and administrative staff. Dr. Chen's research was partially supported by NIH grants #CA 70101 and #CA 74015.

References

Aitkin, M. (1991). Posterior Bayes factor (with discussion). *J. Roy. Statist. Soc. Ser. B* **53**, 111-128.

Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models (with discussion). *J. Amer. Statist. Assoc.* **95**, 1127-1156.

Bayarri, M. J. and Berger, J. O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics* **6** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 53-82. Oxford University Press, Oxford.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition. Springer-Verlag, New York.

Berger, J. O., Brown, L. D. and Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Ann. Statist.* **22**, 1787-1807.

Berger, J. O., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statist. Sci.* **12**, 133-160.

Berger, J. O. and Pericchi, R. L. (2001). Objective Bayesian methods for model selection: introduction and comparison. *Model Selection* (P. Lahiri Editor), Institute of Mathematical Statistics Lecture Notes- Monograph Series **38**.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.* **82**, 112-122.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143**, 383-430.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Second edition. Chapman and Hall, London.

Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25**, 1563-1594.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270-281.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximation. *J. Amer. Statist. Assoc.* **92**, 903-915.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1-13.

Geisser, S., and Eddy, W. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153-160.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163-185.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). Criterion based methods for Bayesian model assessment. *Statist. Sinica* **11**, 419-443.

Jeffreys, H. (1961). *Theory of Probability.* Third edition. Clarendon Press, Oxford.

Kass, R. E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. Roy. Statist. Soc. Ser. B* **54**, 129-144.

Kass, R. E. and Raftery, A. E. (1995). Bayes factor. *J. Amer. Statist. Assoc.* **90**, 773-795.

Lavine, M. and Schervish, M. (1999). Bayes factors: what they are and what they are not. *Amer. Statist.* **53**, 119-122.

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22**, 1142-1160.

Meng, X.-L. and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831-860.

O'Hagan, Q.(1994). *Kendall's Advanced Theory of Statistics.* Volume 2B: Bayesian Inference. Edward Arnold, London.

Vlachos, P. K. and Gelfand, A. E. (2003). On the calibration of Bayesian model choice criteria. *J. Statist. Planning and Inference* **111**, 223-234.

Williams, E. (1959). *Regression Analysis.* Wiley, New York.

Departamento de Economía y Empresa, Universidad de Castilla-La Mancha, Plaza de la Universidad, 1, Albacete, 02071, Spain.

E-mail: Gonzalo.GarciaDonato@uclm.es

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120, U.S.A.

E-mail: mhchen@stat.uconn.edu

(Received May 2003; accepted January 2004)