Nonsubjective Bayes testing—an overview

Jayanta K. Ghosh[a], Tapas Samanta[b,*]

[a]*Stat-Math Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India*
[b]*Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India*

Abstract

In Bayesian model selection, or hypothesis testing, difficulties arise when improper noninformative priors are used to calculate the Bayes factors. Several methods have been proposed to remove these difficulties. In this paper we discuss a unified derivation of some of these methods which shows that in some qualitative or conceptual sense, these methods are no more than a fixed number of observations away from a Bayes factor based on noninformative priors, and are close to each other and to certain Bayes factors based on low information proper priors which include priors recommended by Jeffreys (1961).

MSC: 62A01; 62F15

Keywords: Bayes factor; Model selection; Noninformative prior; Training sample

1. Introduction

Bayesian testing of sharp hypotheses, or model selection, requires specification of prior distributions for the parameters of the proposed models. If a nonsubjective (automatic) viewpoint is adopted, one is motivated to use standard (improper) noninformative priors for the parameters. For a discussion on the need for Bayesian methods in hypothesis testing and the desirability of an automatic method see, for example, Berger and Pericchi (1996a, b) and the references therein. However, there are difficulties with noninformative priors that are improper, and are hence defined only upto arbitrary constant multipliers. The usual Bayes factor, on which a test is based, is thus indeterminate if improper priors are used. A number of methods have been proposed to remove this indeterminacy. In the present paper we discuss a unified derivation of some of these methods. The different nonsubjective methods are found to be close to each other at least in the examples considered in this paper. We try to show in some sense that the nonsubjective Bayes factors may be thought of as an adjustment to a Bayes factor based on noninformative priors, and in some qualitative or conceptual sense, they

are close to each other and to Bayes factors based on low information proper priors which include priors recommended for these problems by Jeffreys (1961). Berger and Pericchi (1996a) feel these automatic methods are one way of generating the kind of proper priors that Jeffreys (1961) recommends as appropriate for testing sharp scientific hypotheses. Several Bayesians do not consider testing a sharp hypothesis a well posed problem. On the other hand Jeffreys (1961) and Edwards et al. (1963) regard it as legitimate object of study. In a concluding section we discuss whether replacement of an improper prior by a proper prior or replacement of a sharp hypothesis by a suitable interval would solve satisfactorily the problem of indeterminacy of the Bayes factor based on improper priors. We also discuss there briefly an alternative nonsubjective method due to Bernardo (1999).

Our main concern is to throw some light on new controversial methods, some of which have been used in our work of automatic geological mapping (Ghosh et al., 1997).

2. Nonsubjective methods of hypothesis testing

2.1. Bayes factor based on noninformative prior

We consider two models M_1 and M_2 for data $\boldsymbol{X}$ with density $f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$ under model M_i, $\boldsymbol{\theta}_i$ being an unknown parameter of dimension p_i, $i=1,2$. Given prior specifications $\pi_i(\boldsymbol{\theta}_i)$ for parameter $\boldsymbol{\theta}_i$ and prior probabilities $P_i(M_i)$ for model M_i, Bayesian hypothesis testing, or model selection is achieved by comparing the posterior probabilities $P(M_i|\boldsymbol{x})$, and hence may be based on the ratio

$$\frac{P(M_2|\boldsymbol{x})}{P(M_1|\boldsymbol{x})}=\frac{P(M_2)}{P(M_1)}B_{21}(\boldsymbol{x}), \tag{1}$$

where $B_{21}=B_{21}(\boldsymbol{x})$, known as the Bayes factor (BF) of M_2 to M_1, is defined by

$$B_{21}=\frac{m_2(\boldsymbol{x})}{m_1(\boldsymbol{x})}=\frac{\int f_2(\boldsymbol{x}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2}{\int f_1(\boldsymbol{x}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1} \tag{2}$$

here $m_i(\boldsymbol{x})$ is the marginal density of $\boldsymbol{X}$ under M_i. When the models are a priori judged equally likely, $P(M_1)=P(M_2)$, and the ratio in (1) is equal to the Bayes factor B_{21}.

In order to compute the Bayes factor B_{21}, the prior distributions $\pi_i(\boldsymbol{\theta}_i)$ need to be specified. Here we look for an automatic (nonsubjective) method of model selection that uses standard (default) noninformative priors.

There is, however, a difficulty with (2) for improper noninformative priors π_i since these are defined only upto arbitrary multiplicative constants c_i and hence B_{21} is determined only upto an arbitrary multiplicative constant c_2/c_1; $c_i\pi_i$ has same properties as π_i implying $(c_2/c_1)B_{21}$ has as much validity as B_{21}. This indeterminacy, noted by Jeffreys (1961), has been the main motivation of new nonsubjective methods. Below, we shall confine mainly to the nested case where f_1 and f_2 are of the same

functional form and $f_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$ is the same as $f_2(\boldsymbol{x}|\boldsymbol{\theta}_2)$ with some of the co-ordinates of $\boldsymbol{\theta}_2$ specified.

It may be mentioned here that the truncation of noninformative priors leads to a large penalty for the more complex model M_2. In Example 1 of Section 2.2 if one uses a uniform prior over $-K \leqslant \mu \leqslant K$, then for large K, the new BF is approximately $1/2K$ times the BF for the noninformative prior with $c_2 = 1$. This is reminiscent of the phenomenon observed by Lindley (1957). A similar conclusion is obtained if one uses as priors $N(0, \tau^2)$, vide Bernardo (1999).

De Finetti's justification of Bayesian analysis and subjective priors through coherence depends on finitely additive proper priors (see, for example, a treatment in Schervish, 1995). One might hope a solution to our problem lies in turning to finitely additive priors. Heath and Sudderth (1978) have shown that in some cases the posterior for an improper countably additive prior can be shown to be the posterior for a proper finitely additive prior. For example, this is the case if in Example 1 (Section 2.2) we consider the posterior for M_2 corresponding to the Lebesgue measure as prior. So it is natural to ask if the posterior for M_1 can be defined in the situation when the improper prior under M_2 is replaced by a proper finitely additive prior with the same posterior as that corresponding to the Lebesgue prior. It may be pointed out that in general posteriors do not exist for finitely additive priors and there are no standard definitions. Heath and Sudderth (1978) define it in the same way as for countably additive priors and Regazzini (1987) defines it in a different way. It turns out that in Example 1, the posterior for M_1 and hence the BF cannot be defined in the sense of Heath and Sudderth (1978), and they have no unique value in the sense of Regazzini (1987) (Ghosh and Ramamoorthi, 2000).

The new nonsubjective methods try to adjust for the arbitrariness in B_{21} that arises due to the arbitrariness of the multiplicative constants c_i of π_i. Several solutions have been proposed in Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), Berger and Pericchi (1996a), O'Hagan (1995) and others, e.g., Pérez (1998) and Iwaki (1997), including Jeffreys (1961). Different alternative nonsubjective methods have been proposed by Bernardo (1999) and Goutis and Robert (1998); see also Bernardo (1980) and Bernardo and Bayarri (1985).

2.2. *Imaginary minimal sample device and related methods*

We assume suitable noninformative priors have been chosen and the only problem is with the constants c_1 and c_2. One way to resolve the indeterminacy of the Bayes factor B_{21} described above is to properly assign a particular value to the (arbitrary) constant multiplier $c = c_2/c_1$. This is achieved if we can imagine a minimal data set $\boldsymbol{x}_0$ and assign a particular value to $B_{21}(\boldsymbol{x}_0)$.

Definition 1. A minimal training sample is a sample with the smallest sample size for which the marginals $m_1(\boldsymbol{x})$ and $m_2(\boldsymbol{x})$ are finite for all $\boldsymbol{x}$.

Spiegelhalter and Smith (1982) propose the following solution for nested models. Imagine a minimal training sample that provides maximum possible support for M_1. In the context of Example 1 below, introspect in an informal way about an imaginary minimal sample of size one with $x_0 = 0$ where one would not like to reject M_1. The idea seems to be that given such a small sample, namely of size one, and the data $x_0 = 0$ being as consistent as possible with M_1, one should not have a preference for either M_1 or M_2. This is ensured by setting $cB_{21}(x_0) = 1$, which in turn resolves the indeterminacy of B_{21} for any sample size and any data. Spiegelhalter and Smith (1982) consider comparison of two nested linear models, generating an imaginary minimal training sample, which leads to an F-statistic value $F = 0$. Here we propose the following solution. Find a minimal sample size and then x_0 for which $m_1(x_0) = \sup_x m_1(x)$ providing maximal support for the simpler model M_1. Then set the adjusted Bayes factor equal to one, i.e.,

$$1 = cB_{21}(x_0) = c\frac{m_2(x_0)}{m_1(x_0)} \tag{3}$$

from which the constant c can be found. The resulting Bayes factor, which we call SSBF_{21} is then obtained as

$$\mathrm{SSBF}_{21} = cB_{21}. \tag{4}$$

An alternative formal way is to determine $c = c_2/c_1$ by solving

$$c_1 \sup_x m_1(x) = c_2 \sup_x m_2(x) \tag{5}$$

where supremum on both sides are over all minimal samples x for which the marginals $m_i(x)$ are finite. This is similar but not identical to (3) since the supremum for m_1 and m_2 may not be obtained at the same x_0. This choice of assigning c amounts to saying that maximum support for both the models are same for this sample size. Since m_2 or both m_1 and m_2 may not integrate to one, (5) is trying to bring m_1 and m_2 to the same scale through a comparison of their suprema.

Example 1 (*Testing normal mean*). Let $X = (X_1, X_2, \ldots, X_n)$. Under M_1, X_i are iid $N(0,1)$ and under M_2, X_i are iid $N(\mu,1)$; $\mu \in \mathbb{R}$ is the unknown mean. Consider the uniform noninformative prior $\pi(\mu) \equiv 1$ for μ. Here $m_1(x)$ and $m_2(x)$ are finite for all x for a sample of size one or more and therefore, by definition, an imaginary minimal training sample x is of size 1 with $m_1(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ and $m_2(x) = 1$. Thus $x_0 = 0$ maximizes $m_1(x)$, and (3) gives $c = c_2/c_1 = 1/\sqrt{2\pi}$. The alternative way of determining c described in (5) above also gives the same answer.

There are, however, examples where both the above options of determining the constant c fail.

Example 2 (*Testing normal mean with variance unknown*). Let $X = (X_1, \ldots, X_n)$.

M_1: $X_1, \ldots, X_n$ are iid $N(0, \sigma_1^2)$,

M_2: $X_1, \ldots, X_n$ are iid $N(\mu, \sigma_2^2)$.

Consider the noninformative priors $\pi_1(\sigma_1)=1/\sigma_1$ and $\pi_2(\mu,\sigma_2)=1/\sigma_2$. Here, an imaginary minimal sample is of size 2 and for such a sample (x, y),

$$m_1(x, y)=\frac{1}{2\pi(x^2+y^2)} \quad \text{and} \quad m_2(x, y)=\frac{1}{2|x-y|}. \tag{6}$$

Thus $\sup m_i(x, y)=\infty$ and the above methods do not apply.

We present below a version of the above method that works more generally. Consider first the particular case of Example 1. Let $\bar{x}$ be the observed mean. The observed Fisher information (per unit observation) here is $\hat{I}=1$. Consider a (data dependent) uniform prior for μ over $\bar{x}\mp k/\sqrt{\hat{I}}$, i.e., over $(\bar{x}-k,\ \bar{x}+k)$ for a suitable constant k. Such data dependent priors are in the spirit of parametric empirical Bayesian inference where some hyperparameters in the last layer of a hierarchical prior are estimated from data, see, for example, Morris (1983). It is known that the estimates obtained in this way are close to the proper Bayes estimates but the posterior uncertainty in these estimates is neglected in the empirical Bayes analysis. It appears that in many cases the new data dependent priors can be thought of as approximations to proper priors. See in this connection the discussion of intrinsic priors in Section 2.5 and also the concluding section.

The Bayes factor corresponding to the data dependent uniform prior over $(\bar{x}-k,\ \bar{x}+k)$ in Example 1 turns out to be $(1/2k)[\Phi(k\sqrt{n})-\Phi(-k\sqrt{n})]B_{21}$ where B_{21} is the Bayes factor with the uniform noninformative prior for μ and Φ is the cdf of $N(0, 1)$. Thus using such a data dependent proper uniform prior corresponds to having $c_2/c_1=(1/2k)[\Phi(k\sqrt{n})-\Phi(-k\sqrt{n})]$. In other words, the Spiegelhalter–Smith method of choosing c_2/c_1 is equivalent to considering a uniform prior over an interval of the form $\bar{x}\mp k$ for suitable k.

Consider now the general case. Let $\hat{\boldsymbol{\theta}}_i$ be the MLE of $\boldsymbol{\theta}_i$ and $\hat{I}_{\text{in}}=(-\partial^2\log f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)/\partial\theta_{ik}\partial\theta_{il})|_{\hat{\boldsymbol{\theta}}}$ be the observed Fisher information matrix under M_i. Based on the reciprocal of this, choose an ellipsoid around the MLE $\hat{\boldsymbol{\theta}}_i$ that contains $\boldsymbol{\theta}_i$ with (approximate) probability $(1-\alpha)$ for some suitably chosen small α, $0<\alpha<1$ (e.g., $\alpha=0.05$). If the noninformative prior π_i is (improper) uniform, choose a data dependent prior $\pi_i^*(\boldsymbol{\theta}_i)$ that is uniform over $\sqrt{n/m}$ times the region enclosed by this ellipsoid around $\hat{\boldsymbol{\theta}}_i$, where m is the size of a minimal sample. Under regularity conditions $\hat{I}_{\text{in}}^{1/2}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)$ is $\text{AN}(\mathbf{o}, I_{p_i})$ where I_p is the identity matrix of order p. So, the ellipsoid around $\hat{\boldsymbol{\theta}}_i$ that contains $\boldsymbol{\theta}_i$ with approximate probability $(1-\alpha)$ is

$$\{\boldsymbol{\theta}_i:\ (\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)'\hat{I}_{\text{in}}(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)\leqslant\chi^2_{p_i,\alpha}\},$$

where $\chi^2_{p_i,\alpha}$ is the upper α-point of χ^2 distribution with p_i degrees of freedom. Thus we take $\pi_i^*(\boldsymbol{\theta}_i)$ to be uniform over the region

$$\left\{\boldsymbol{\theta}_i:\ (\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)'\hat{I}_{\text{in}}(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)\leqslant\frac{n}{m}\chi^2_{p_i,\alpha}\right\}.$$

If the noninformative prior π_i is not uniform, π_i^* is taken to be π_i truncated on the above region. We now use these data dependent priors $\pi_1^*(\boldsymbol{\theta}_1)$ and $\pi_2^*(\boldsymbol{\theta}_2)$ for models

M_1 and M_2 to calculate the Bayes factor which we call DUBF (BF based on data dependent uniform prior). Note that each π_i^* is diffuse but "not in conflict with data" since π_i^* is centered around $\hat{\theta}_i$. In Section 2.5 we shall exhibit similar data dependent priors for other adjustments of the Bayes factors proposed by Berger and Pericchi (1996a) and O'Hagan (1995). They can be thought of as some sort of approximations to proper priors (see Section 2.5).

In the nested case when it is reasonable to consider the same noninformative prior for the "shared" part of the parameter (such as in the case with location-scale family like Example 2), we propose another way to choose the data dependent prior. Suppose that we can write $\theta_2 = (\theta_1, \eta)$ where η is the extra parameter that is specified under M_1, the shared parameter θ_1 has the same interpretation under both the models, and the same noninformative prior is considered for θ_1 under these models. Let $\pi_2(\theta_2) = \pi_1(\theta_1)\pi_2(\eta|\theta_1)$. We propose the following. Use $\pi_1(\theta_1)$ itself for θ_1 in both the models. For η (conditioned on θ_1) choose a data dependent prior concentrated on the region enclosed by the ellipsoid centered at the MLE of η given θ_1 as described in the preceding paragraph.

Example 2 (*continued*). For model M_1, $p_1 = 1$, $\hat{\sigma}_1 = \sqrt{(1/n)\sum x_i^2}$, $\hat{I}_{1n} = 2n/\hat{\sigma}_1^2$ and π_1^* is taken as the prior $\pi_1(\sigma_1) = 1/\sigma_1$ truncated on the interval $(\max(\hat{\sigma}_1 - \frac{1}{2}z_{\alpha/2}\hat{\sigma}_1, 0), \hat{\sigma}_1 + \frac{1}{2}z_{\alpha/2}\hat{\sigma}_1)$ where $z_{\alpha/2}$ is the upper $\alpha/2$-point of $N(0,1)$.

For M_2, $p_2 = 2$, $\hat{\mu} = \bar{x}$, $\hat{\sigma}_2 = \sqrt{(1/n)\sum (x_i - \bar{x})^2}$, $\hat{I}_{2n} = (n/\hat{\sigma}_2^2)\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and π_2^* is $\pi_2(\mu, \sigma_2) = 1/\sigma_2$ truncated on the region

$$\{(\mu, \sigma_2) \in \mathbb{R} \times \mathbb{R}^+ : (\mu - \hat{\mu})^2 + 2(\sigma_2 - \hat{\sigma}_2)^2 \leqslant \tfrac{1}{2}\chi_{2,\alpha}^2 \hat{\sigma}_2^2\}.$$

The other option suggests choosing $\pi_1^*(\sigma_1) = 1/\sigma_1$, $\pi_2^*(\mu, \sigma_2) = (1/\sigma_2)\pi_2^*(\mu|\sigma_2)$ where $\pi_2^*(\mu|\sigma_2)$ is uniform over the interval $\bar{x} \mp z_{\alpha/2}\sigma_2/\sqrt{2}$; here $\hat{\mu} = \bar{x}$, $\hat{I}_n = n/\sigma_2^2$ and $m = 2$. The resulting DUBF (BF with priors π_1^* and π_2^*) is given by

$$\frac{1}{2k}[\Phi(k\sqrt{n}) - \Phi(-k\sqrt{n})]\sqrt{\frac{2\pi}{n}}\left(1 + \frac{t^2}{n-1}\right)^{n/2},$$

where $t = \sqrt{n}\bar{x}/s$ is the t-statistic and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$. Thus, DUBF is a function of t only and goes to ∞ as $|t|$ goes to ∞.

Calculations, with several data sets, of SSBF or DUBF and other nonsubjective Bayes factors described in the following sections for Examples 1 and 2 indicate that these BFs are close to each other and they tend to cluster away from the likelihood ratio or the BF based on noninformative priors. All these Bayes factors depend on data dependent priors but they seem to depend on data in apparently different ways. So it is somewhat surprising that they should be close. We do not have any proof of this, nor do we know if this phenomenon holds for non-normal examples also.

2.3. The intrinsic Bayes factor

Another solution to the problem with improper priors is to use part of the data as a *training sample*. The idea is to use the training sample to obtain proper posterior distributions for the parameters which can then be used as priors to compute a Bayes factor with the remainder of the data.

Let $x_1, x_2, \ldots, x_n$ constitute the whole sample. For a subsample $x_{j_1}, x_{j_2}, \ldots, x_{j_m}$ $(1 \leqslant j_1 < j_2 < \cdots < j_m \leqslant n)$, the posterior density of θ_i given $x_{j_1}, \ldots, x_{j_m}$ under M_i is given by

$$\pi_i(\theta_i|x_{j_1}, \ldots, x_{j_m}) = \frac{f_i(x_{j_1}, \ldots, x_{j_m}|\theta_i)\pi_i(\theta_i)}{m_i(x_{j_1}, \ldots, x_{j_m})}$$

$$= \frac{f_i(x_{j_1}, \ldots, x_{j_m}|\theta_i)\pi_i(\theta_i)}{\int f_i(x_{j_1}, \ldots, x_{j_m}|\theta_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i}, \quad i = 1, 2. \tag{7}$$

Berger and Pericchi (1996a) use training sample of minimal size, leaving most part of the data for model comparison. Let m be the minimum sample size such that $\pi_i(\theta_i|x_{j_1}, \ldots, x_{j_m})$, $i = 1, 2$, are proper or equivalently, $m_i(x_{j_1}, \ldots, x_{j_m})$, $i = 1, 2$, are finite. Let $x_{j_1}, \ldots, x_{j_m}$ be such a minimal training sample. The Bayes factor with the remainder of the data using the above $\pi_i(\theta_i|x_{j_1}, \ldots, x_{j_m})$ in (7) as priors (conditional BF) is given by

$$B_{21}(j_1, \ldots, j_m) = \frac{\int (f_2(x_1, \ldots, x_n|\theta_2)/f_2(x_{j_1}, \ldots, x_{j_m}|\theta_2))\pi_2(\theta_2|x_{j_1}, \ldots, x_{j_m})\,\mathrm{d}\theta_2}{\int (f_1(x_1, \ldots, x_n|\theta_1)/f_1(x_{j_1}, \ldots, x_{j_m}|\theta_1))\pi_1(\theta_1|x_{j_1}, \ldots, x_{j_m})\,\mathrm{d}\theta_1}$$

$$= B_{21}\frac{m_1(x_{j_1}, \ldots, x_{j_m})}{m_2(x_{j_1}, \ldots, x_{j_m})}. \tag{8}$$

It is to be noted that the arbitrary constant multiplier c_2/c_1 of B_{21} is cancelled by that (c_1/c_2) of $m_1(x_{j_1}, \ldots, x_{j_m})/m_2(x_{j_1}, \ldots, x_{j_m})$ so that the indeterminacy of the Bayes factor is removed in (8). However, this conditional BF in (8) depends on the choice of the training sample $x_{j_1}, \ldots, x_{j_m}$. Berger and Pericchi (1996a) suggest considering all possible training samples and taking average of the $\binom{n}{m}$ conditional BFs $B_{21}(j_1, \ldots, j_m)$'s to obtain what is called the intrinsic Bayes factor (IBF). For example, taking an arithmetic average leads to

$$\mathrm{AIBF}_{21} = B_{21}\frac{1}{\binom{n}{m}}\sum\frac{m_1(x_{j_1}, \ldots, x_{j_m})}{m_2(x_{j_1}, \ldots, x_{j_m})} \tag{9}$$

while the geometric average gives

$$\mathrm{GIBF}_{21} = B_{21}\left(\prod\frac{m_1(x_{j_1}, \ldots, x_{j_m})}{m_2(x_{j_1}, \ldots, x_{j_m})}\right)^{1/\binom{n}{m}}, \tag{10}$$

the sum and product in (9) and (10) being taken over the $\binom{n}{m}$ possible training samples $x_{j_1}, \ldots, x_{j_m}$ with $1 \leqslant j_1 < \cdots < j_m \leqslant n$.

Berger and Pericchi (1996a) also suggest using trimmed averages or the median (complete trimming) of the conditional BFs when taking an average of all the conditional BFs does not seem reasonable (e.g., when the conditional BFs vary much). AIBF

and GIBF have good properties but are affected by outliers. If the sample size is very small, using a part of the sample as a training sample may be impractical and Berger and Pericchi (1996a) recommend using expected intrinsic Bayes factors that replace the averages in (9) and (10) by their expectations, evaluated at the MLE under M_2. The AIBF is justified by the possibility of its correspondence to actual Bayes factors with respect to"intrinsic" proper priors at least asymptotically.

The idea of a training sample has also been used by many others including Lempers (1971), Atkinson (1978), Geisser and Eddy (1979), Spiegelhalter and Smith (1982), San Martini and Spezzaferri (1984) and Gelfand et al. (1992).

2.4. The fractional Bayes factor

O'Hagan (1995) proposes a solution using a fractional part of the full likelihood in place of using training samples and averaging over them. The resulting "partial" Bayes factor, called the fractional Bayes factor (FBF) is given by

$$\mathrm{FBF}_{21} = \frac{m_2(\boldsymbol{x}, b)}{m_1(\boldsymbol{x}, b)}, \tag{11}$$

where b is a fraction and

$$m_i(\boldsymbol{x}, b) = \frac{\int f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\,\mathrm{d}\boldsymbol{\theta}_i}{\int [f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)]^b \pi_i(\boldsymbol{\theta}_i)\,\mathrm{d}\boldsymbol{\theta}_i}. \tag{12}$$

To make the FBF comparable with the IBF we take $b = m/n$ where m is the size of a minimal training sample as defined in the case of IBF or SSBF. O'Hagan recommends other choices of b also (viz. $b = \sqrt{n}/n$ or $\log n/n$) but we ignore these in the present paper.

2.5. Data dependent priors

We now introduce a data dependent prior π_i^* which is easy to "calibrate" with respect to the noninformative prior π_i. Similar but more general priors have also been introduced by Pérez (1998) and studied in detail for linear models and mixtures.

Let $\pi_i(\boldsymbol{\theta}_i|x_{j_1}, \ldots, x_{j_m})$ be the posterior given the training sample $x_{j_1}, \ldots, x_{j_m}$, $1 \leqslant j_1 < \cdots < j_m \leqslant n$ under model M_i as given in (7). We start with $\pi_i(\boldsymbol{\theta}_i|x_{j_1}, \ldots, x_{j_m})$ as prior that differs from the noninformative prior $\pi_i(\boldsymbol{\theta}_i)$ by m observations (in the same sense we calibrate conjugate priors). Consider the arithmetic average of these priors

$$\pi_i^*(\boldsymbol{\theta}_i) = \frac{1}{\binom{n}{m}} \sum_{j_1 < \cdots < j_m} \pi_i(\boldsymbol{\theta}_i|x_{j_1}, \ldots, x_{j_m}) \quad (i = 1, 2) \tag{13}$$

which is more diffuse than each $\pi_i(\boldsymbol{\theta}_i|x_{j_1}, \ldots, x_{j_m})$ and so presumably differs from the noninformative priors by "no more than m observations". If we construct a Bayes factor using π_i^* in place of π_i, we get a new "calibrated" BF.

We now define another data dependent prior by taking the geometric mean

$$\pi^*_{i,gm}(\theta_i) = \left[\prod_{1 \leqslant j_1 < \cdots < j_m \leqslant n} \pi_i(\theta_i | x_{j_1}, \ldots, x_{j_m}) \right]^{1/\binom{n}{m}}, \quad i = 1, 2. \tag{14}$$

By the same analogy as before this new $\pi^*_{i,gm}$ is also calibrated and is "no more than m observations away from π_i". Unfortunately, $\pi^*_{i,gm}$ is not a probability but let us ignore this for the time being.

If we calculate the BF using this data dependent prior we will be using the same data more than once. An adjustment is called for. Let us consider iid observations. Since number of training samples $\{x_{j_1}, \ldots, x_{j_m}\}$ that include a particular observation x_j is $\binom{n-1}{m-1}$, noting that $\binom{n-1}{m-1}/\binom{n}{m} = m/n$, an adjusted likelihood is

$$\left[\prod_{j=1}^{n} f_i(x_j) \right]^{1-m/n}.$$

Let $B_{21,gm}$ be the adjusted Bayes factor based on this adjusted likelihood and the data dependent priors $\pi^*_{i,gm}$. It turns out that

Result 1. The adjusted BF $B_{21,gm}$ and GIBF_{21} are identical.

To see this note that

$$\pi^*_{i,gm}(\theta_i) = \frac{[f_i(x_1, \ldots, x_n | \theta_i)]^{m/n} \pi_i(\theta_i)}{(\prod_{j_1 < \cdots < j_m} m_i(x_{j_1}, \ldots, x_{j_m}))^{1/\binom{n}{m}}}.$$

Therefore

$$\begin{aligned} B_{21,gm} &= \frac{\int [f_2(x_1, \ldots, x_n)]^{1-m/n} \pi^*_{2,gm}(\theta_2)\, \mathrm{d}\theta_2}{\int [f_1(x_1, \ldots, x_n)]^{1-m/n} \pi^*_{1,gm}(\theta_1)\, \mathrm{d}\theta_1} \\ &= B_{21} \left(\prod_{j_1 < \cdots < j_m} \frac{m_1(x_{j_1}, \ldots, x_{j_m})}{m_2(x_{j_1}, \ldots, x_{j_m})} \right)^{1/\binom{n}{m}} \\ &= \mathrm{GIBF}_{21}. \end{aligned}$$

Thus the GIBF, and by analogy the other IBFs also "do not differ from the Bayes factor based on noninformative priors by more than m observations".

One way of adjusting for the fact that $\pi^*_{i,gm}$ is not a probability is to first take a geometric mean

$$\left[\prod_{j_1 < \cdots < j_m} \pi_i(\theta_i) f_i(x_{j_1}, \ldots, x_{j_m} | \theta_i) \right]^{1/\binom{n}{m}}$$

and then normalize to one. Let us denote this adjusted data dependent prior by $\pi^*_{i,agm}$. Note that

$$\pi^*_{i,agm}(\theta_i) = \frac{[f_i(\boldsymbol{x} | \theta_i)]^{m/n} \pi_i(\theta_i)}{\int [f_i(\boldsymbol{x} | \theta_i)]^{m/n} \pi_i(\theta_i)\, \mathrm{d}\theta_i}.$$

Suppose we use $\pi^*_{L,agjm}$ and the same adjusted likelihood as before to calculate the Bayes factor which we call $B_{21,agjm}$. Then we have the following.

Result 2. FBF_{21} and the adjusted BF $B_{21,agjm}$ are identical.

If we want to use the arithmetic mean data dependent prior π^*_i defined in (13) to calculate the BF, an adjustment of the likelihood is to be made. An adjusted likelihood in this case is

$$\sum_{j_1<\cdots<j_m} \frac{f_i(x_1,\ldots,x_n|\theta_i)}{f_i(x_{j_1},\ldots,x_{j_m}|\theta_i)} w^{(i)}_{j_1\cdots j_m},$$

where

$$w^{(i)}_{j_1\cdots j_m} = \frac{\pi_i(\theta_i|x_{j_1},\ldots,x_{j_m})}{\binom{n}{m}\,\pi^*_i(\theta_i)} = \frac{\pi_i(\theta_i|x_{j_1},\ldots,x_{j_m})}{\sum_{j_1<\cdots<j_m}\pi_i(\theta_i|x_{j_1},\ldots,x_{j_m})}.$$

Let $B_{21,am}$ be the adjusted BF based on this adjusted likelihood and the data dependent priors $\pi^*_i(\theta_i)$. Then we have

$$B_{21,am} = B_{21}\frac{\binom{n}{m}^{-1}\sum_{j_1<\cdots<j_m} 1/m_2(x_{j_1},\ldots,x_{j_m})}{\binom{n}{m}^{-1}\sum_{j_1<\cdots<j_m} 1/m_1(x_{j_1},\ldots,x_{j_m})}$$

$$= B_{21}\tilde{B}_{12} \quad \text{(say)}. \tag{15}$$

It is to be noted that unlike AIBF, this adjusted BF based on data dependent prior π^*_i is "multiple model coherent" (see Berger and Pericchi, 1996a, Section 5) and thus *directly yields "pseudo" posterior probabilities of the considered models.* This is also true for the other adjusted or unadjusted BFs based on data dependent priors.

Example 1 (*continued*). Here $B_{21} = \sqrt{2\pi/n}\exp((n/2)\bar{x}^2)$ and the conditional BFs are given by $B_{21}(x_i) = B_{21}(1/\sqrt{2\pi})\exp(-x_i^2/2)$. The adjusted BF with data dependent prior (13) is

$$B_{21,am} = B_{21}\frac{n}{\sum_{i=1}^n \sqrt{2\pi}\exp(x_i^2/2)}$$

which is the harmonic mean of the conditional BFs while the AIBF and GIBF are given by the arithmetic and geometric means respectively.

Berger and Pericchi (1996a) justify their Bayes factors by the possibility of the existence of "intrinsic" priors. Their arguments may work with $B_{21,am}$, replacing their correction factor (the multiplier of B_{21} in (9)) by $\tilde{B}_{12}$ of (15) above. As suggested by Berger and Pericchi, a solution to the intrinsic prior determining equations corresponding to $B_{21,am}$ would be given by

$$\pi^I_1(\theta_1) = \pi_1(\theta_1), \quad \pi^I_2(\theta_2) = \pi_2(\theta_2)B^*(\theta_2),$$

where π^I_i are the intrinsic priors and $B^*(\theta_2)$ is the limit of $\tilde{B}_{12}$ under M_2 as $n\to\infty$.

Example 2 (*continued*). Here $m=2$, $m_1(x_1,x_2)$ and $m_2(x_1,x_2)$ are as given in (6). Thus

$$B^*(\mu,\sigma_2)=\frac{E|x_1-x_2|}{\pi E(x_1^2+x_2^2)}=\frac{\sigma_2}{\pi\sqrt{\pi}(\sigma_2^2+\mu^2)}$$

and thus the intrinsic priors corresponding to $B_{21,am}$ are

$$\pi_1^I(\sigma_1)=\frac{1}{\sigma_1},\qquad \pi_2^I(\mu,\sigma_2)=\frac{1}{\sigma_2}\pi_2^I(\mu|\sigma_2)=\frac{1}{\sigma_2}\frac{1}{\sqrt{\pi}},\ \text{Cauchy } (0,\sigma_2) \text{ prior for } \mu. \tag{16}$$

Note that $\pi_2^I(\mu|\sigma_2)$ here is a constant multiple of Jeffreys's Cauchy $(0,\sigma_2)$ choice of $\pi_2(\mu|\sigma_2)$.

The intrinsic priors suggested by Berger and Pericchi corresponding to the AIBF (9) turns out to be

$$\pi_1^I(\sigma_1)=\frac{1}{\sigma_1},\qquad \pi_2^I(\mu,\sigma_2)=\frac{1}{\sigma_2}\pi_2^I(\mu|\sigma_2)$$

$$\text{with } \pi_2^I(\mu|\sigma_2)=\frac{1}{\pi\sqrt{\pi}\sigma_2}\exp(-\mu^2/\sigma_2^2)\sum_{r=0}^{\infty}\frac{(\mu^2/\sigma_2^2)^r}{\Gamma(r+1)(r+1/2)}. \tag{17}$$

It is to be noted that $\int_{-\infty}^{\infty}\pi_2^I(\mu|\sigma_2)\,\mathrm{d}\mu=1$.

Berger and Pericchi (1996a) consider the noninformative priors $\pi_1=1/\sigma_1$ and $\pi_2(\mu,\sigma_2)=1/\sigma_2^2$ for computational simplicity and obtain

$$\pi_2^I(\mu,\sigma_2)=\frac{1}{\sigma_2}\pi_2^I(\mu|\sigma_2)=\frac{1}{\sigma_2}\frac{1-\exp(-\mu^2/\sigma_2^2)}{2\sqrt{\pi}(\mu^2/\sigma_2)}, \tag{18}$$

where $\pi_2^I(\mu|\sigma_2)$ is a proper prior close to the Cauchy $(0,\sigma_2)$ prior for μ. If we use these noninformative priors, the intrinsic priors corresponding to $B_{21,am}$ are obtained as

$$\pi_1^I(\sigma_1)=\frac{1}{\sigma_1},\qquad \pi_2^I(\mu,\sigma_2)=\frac{1}{\sigma_2}\pi_2^I(\mu|\sigma_2)=\frac{1}{\sigma_2}\frac{\sqrt{\pi}}{2},\ \text{Cauchy } (0,\sigma_2) \text{ prior for } \mu. \tag{19}$$

3. Examples

In this section we present a few examples illustrating points on various aspects of the nonsubjective BFs described in Section 2.

Example 3 (*Nonregular case*). Berger (1997) gave the following example of nonregular case where FBF does not perform well. Let the observations $x_1,\ldots,x_n$ be iid with a common density $f(x,\theta)=\mathrm{e}^{-(x-\theta)}$, $x>\theta$. Consider the problem of comparing the models M_1: $\theta=0$ and M_2: $\theta>0$ with the prior $\pi(\theta)=1$ on $\theta>0$. Then FBF (with fraction b) is given by

$$\mathrm{FBF}=\frac{\int_0^{x_{(1)}}\mathrm{e}^{n\theta}\,\mathrm{d}\theta}{\int_0^{x_{(1)}}\mathrm{e}^{nb\theta}\,\mathrm{d}\theta}=\frac{b(\mathrm{e}^{nx_{(1)}}-1)}{\mathrm{e}^{bnx_{(1)}}-1}, \tag{20}$$

where $x_{(1)} = \min(x_1, \ldots, x_n)$. Clearly, FBF $\geqslant 1$ for any $0 < b < 1$ and any data. The AIBF, on the other hand, performs well for this nonregular example.

We now reexamine this nonregular example. As mentioned above we take $b = 1/n$, size of a minimal training sample being 1. Noting that $x_{(1)}$, being a sufficient statistic, contains too much information about θ, we use a fractional part (with $b = 1/(n-1)$) of the joint likelihood based on the observations other than $x_{(1)}$ in stead of using a fractional part (with $b = 1/n$) of the joint likelihood based on all the n observations to find an FBF. Thus a modified version of the FBF is obtained as

$$\mathrm{FBF}^* = \frac{e^{nx_{(1)}} - 1}{(n-1)(e^{x_{(2)}} - 1)} \tag{21}$$

$x_{(i)}$ is the ith order statistic in $(x_1, \ldots, x_n)$. Probability of FBF* being less than 1 and 2 under M_1 is then approximately equal to 0.60 and 0.83, respectively for moderate n such as $n \geqslant 10$. Also under M_2, both FBF and FBF* tend to ∞ as $n \to \infty$.

If one is motivated to use as little of the data as necessary for a training sample leaving the rest for model selection, one would use a conditional Bayes factor (CBF) conditioned on $x_{(n)}$, the maximum of the observations. This CBF is given by

$$\mathrm{CBF}^* = \frac{e^{nx_{(1)}} - 1}{e^{x_{(n)}} - 1}. \tag{22}$$

It can be shown that as $n \to \infty$, CBF* tends to 0 under M_1 and to ∞ under M_2.

If $x_{(n)}$ is an outlier one can use a few of the extreme observations such as $x_{(n)}$, $x_{(n-1)}$ and $x_{(n-2)}$ and use (a fractional part of) the joint likelihood of these observations to obtain a "partial" Bayes factor.

Example 4. While the above nonregular example of Berger (1997) illustrates that FBF in its original form behaves badly but IBF performs well, O'Hagan (1995) uses "Darwin's Data", a set of data with outliers (see, e.g., Box and Tiao, 1962) in order to show that FBF performs better than the AIBF with respect to sensitivity to outliers. We refer to O'Hagan (1995, p. 114) for details. However, as mentioned earlier in Section 2.3, in such cases one could use the median IBF which would eliminate the sensitivity to outliers. For a suitable transformation of the Darwin data, our problem is the same as that given in Example 1. O'Hagan (1995) obtains the conditional Bayes factors (CBFs) of M_1 to M_2 using the reciprocal of the expression in (8) and calculates AIBF (of M_1 to M_2) as arithmetic average of these CBFs which is not the same as the reciprocal of the AIBF (of M_2 to M_1) recommended by Berger and Pericchi (1996a). The largest and smallest CBFs are reported as 48 968 and 3.85 and the values of the AIBF and FBF are 3364 and 3.814, respectively. With the definitions of this paper, the AIBF, Median IBF and FBF (of M_2 to M_1) are, respectively, 0.1511, 0.2054 and 0.2622; the SSBF or DUBF proposed in Section 2.2 is 0.2625.

We now examine (through examples) the effect of using training samples that are not minimal.

Example 1 (*continued*). Size of a minimal training sample here is 1; the corresponding intrinsic prior for μ is given by a $N(0,2)$ density. If we use training samples of size 2 then the corresponding "intrinsic" prior would be a $N(0,1)$ distribution. This is more peaked than the usual intrinsic prior.

Example 2 (*continued*). Here minimal training samples consist of 2 observations and the intrinsic priors are as given in (17).

If we use training samples of size 3, the "intrinsic" priors would be given by

$$\pi_1^I(\sigma_1) = \frac{1}{\sigma_1},$$

$$\pi_2^I(\mu,\sigma_2) = \frac{1}{\sigma_2}\pi_2^I(\mu|\sigma_2) = \frac{1}{\sigma_2}\frac{\sqrt{3}}{2\sqrt{2}\sigma_2}\exp(-3\mu^2/2\sigma_2^2)\sum_{r=0}^{\infty}\frac{(3\mu^2/2\sigma_2^2)^r}{\Gamma(r+5/2)}, \tag{23}$$

where $\pi_2^I(\mu|\sigma_2)$ in (23) is a proper prior.

The only difference is in $\pi_2^I(\mu|\sigma_2)$ and one can check (simply by plotting) that $\pi_2^I(\mu|\sigma_2)$ of (23) is more peaked than the normalized $\pi_2^I(\mu|\sigma_2)$ of (17).

If we use the noninformative prior $\pi_2(\mu,\sigma_2) = 1/\sigma_2^2$ as in Berger and Pericchi (1996a), the intrinsic priors for training samples of size 2 are as given in (18) and the only change with training samples of size 3 is in

$$\pi_2^I(\mu|\sigma_2) = \frac{\sqrt{3}}{\sqrt{2\pi}\sigma_2}\sum_{r=0}^{\infty}\exp(-3\mu^2/2\sigma_2^2)\frac{(3\mu^2/2\sigma_2^2)^r}{r!}\frac{\Gamma(r+\frac{3}{2})\Gamma(r/2)}{(r+2)!} \tag{24}$$

which is a proper prior that is found (through plotting) to be more peaked than the usual intrinsic proper prior $\pi_2^I(\mu|\sigma_2)$ of (18).

We have seen through Examples 1 and 2 above that use of larger training samples corresponds to Bayes factors with more peaked (intrinsic) priors. We now give an argument as to why it is expected to be so in the general nested case. Consider iid observations and suppose that we are using training samples of size r. Consider only the case with AIBF. Intrinsic priors suggested by Berger and Pericchi (1996a) are

$$\pi_1^I(\boldsymbol{\theta}_1) = \pi_1(\boldsymbol{\theta}_1) \quad \text{and} \quad \pi_2^I(\boldsymbol{\theta}_2) = \pi_2(\boldsymbol{\theta}_2)B(\boldsymbol{\theta}_2), \tag{25}$$

where

$$\begin{aligned} B(\boldsymbol{\theta}_2) &= E_{\boldsymbol{\theta}_2}^{M_2} B_{12}(x_1,\ldots,x_r) \\ &= E_{\boldsymbol{\theta}_2}^{M_2}\left[\frac{m_1(x_1,\ldots,x_r)}{m_2(x_1,\ldots,x_r)}\right]. \end{aligned} \tag{26}$$

Then, for any measurable $\boldsymbol{\theta}_2$-set A, one can show that

$$\int_A \pi_2^I(\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2 = \int \pi_2(A|x_1,\ldots,x_r)m_1(x_1,\ldots,x_r)\,\mathrm{d}x_1\cdots\mathrm{d}x_r. \tag{27}$$

Here $\pi_2(A|x_1,\ldots,x_r)$ denotes the probability of A under the posterior distribution $\pi_2(\cdot|x_1,\ldots,x_r)$. Since the posterior $\pi_2(\cdot|x_1,\ldots,x_r)$ is expected to be more peaked for larger r, the corresponding intrinsic prior satisfying (27) is also expected to be so.

4. Remarks

This section contains some remarks summarizing major features of nonsubjective Bayes factors. This is followed in the next section by a discussion of some controversial issues relating to them.

1. The nonsubjective Bayes factors are obtained by relatively small adjustments to the Bayes factors based on data dependent priors. They tend to attach more penalty to M_2 than Bayes factors based on noninformative priors but less than the Bayes factors based on proper priors obtained by truncating a noninformative prior. The data dependent priors try to reconcile being diffuse and not in conflict with data.

2. GIBF and FBF (and in a sense the other IBFs) are "no more than m observations away" from the Bayes factors based on noninformative priors.

3. In most examples these nonsubjective Bayes factors are close to each other. So it is tempting to conclude that *they mean something.*

4. There are examples where these Bayes factors may differ a lot. But scrutiny shows which are "right". Usually, there are obvious natural adjustments to the "wrong" ones that bring them close to the "right" ones.

Above we have compared the different nonsubjective methods with respect to fixed noninformative priors for the two models. Conclusions remain more or less the same for other noninformative priors.

5. Discussion

The use of improper or data dependent priors as well as putting a positive prior probability on a sharp hypothesis has come in for criticism from several Bayesians. Moreover, there have been other nonsubjective Bayesian approaches to these problems which are different from the methods discussed earlier in the paper. These aspects are briefly discussed in this section. To fix ideas and for simplicity we will consider Example 1 only.

5.1. Improper priors

Various people, e.g., O' Hagan (1995) and Bernardo (1999) have pointed out that even if one has a proper prior given M_2, the Bayes factor is highly non-robust with respect to the choice of prior. Non-robustness plays the same role as indeterminacy. For example, let BF_{21}^N and BF_{21}^{norm} denote the Bayes factors corresponding to the Lebesgue measure $c\,d\mu$ and the normal prior $N(0,\tau^2)$, respectively. Then

$$BF_{21}^N = \sqrt{2\pi}cn^{-1/2}\exp\left[\frac{1}{2}n\bar{x}^2\right],$$

$$BF_{21}^{\text{norm}} = (\tau^2 + 1/n)^{-1/2}n^{-1/2}\exp\left[\frac{1}{2}\frac{n\tau^2}{n\tau^2+1}n\bar{x}^2\right]$$

which clearly indicates similar behaviour of these Bayes factors and the similar roles of $\sqrt{2\pi}c$ and $(\tau^2+1/n)^{-1/2}$. The situation is very different from the posterior robustness that one observes for estimation problems.

Both Bernardo (1999) and Lindley, in his discussion of Bernardo (1999), have noted the effect of τ^2 on the Bayes factor. Bernardo argues that this makes the Bayes Factor an inappropriate tool. Lindley feels this has to do with many different contexts, but remains positive about using a Bayes Factor. His ideas about choosing τ^2 are different from the ideas in the next section but we see some similarities.

5.2. *Data dependent priors and scaling problem*

We only illustrate with our version of the Spiegelhalter–Smith choice and compare with uniform or normal priors.

Suppose a client brings to us the testing problem of Example 1 along with the data $\bar{x}\neq 0$ and the information that the sample size is n. It may be argued that the client has an intuitive feeling that there is some evidence against M_1 but does not know if the evidence is strong enough. If we think in terms of the client's subconscious prior, this would correspond to $\bar{x}$ being in the support of the prior given M_2. If this were not so, the client wouldn't feel there is some evidence.

If we further assume that the client's prior is, at least approximately, uniform, then

$$\pi(\theta)=k, \quad \text{in some interval } J,$$
$$=0, \quad \text{outside},$$

where k is a constant. Typically, J would contain both zero and $\bar{x}$ and would not be a big interval. For example, there would be little prior expectation for data that deviates a lot from zero, like $\bar{x}=10$. The prior we have in mind is similar to Jeffreys's recommendation of $N(0,\tau^2)$ where τ^2 and the population variance (of X) $\sigma^2=1$ are of the same order of magnitude.

A data dependent prior (π_{SS}) is an approximation to this prior in the sense that BF_{21} with client's prior and BF_{21} with π_{SS} will not differ much for *the kind of data considered above*. The approximation would be very poor if $\bar{x}$ is actually 10 but then a client will not come to a statistician. This is a tentative belief based on our own experience; we have never seen such large deviations in the σ-scale.

The scale of the prior is very important. According to classical (frequentist) statisticians, whether $\bar{x}$ is large or not should be measured in the $\sigma/\sqrt{n}$ scale, i.e., by looking at $\sqrt{n}\bar{x}/\sigma$. For a Bayesian this would amount to taking a prior of a comparable scale and with support containing $\bar{x}$. Of course the inference based on Bayes factors with such priors would be very different from Bayes factors with priors having scales of the same order of magnitude as σ.

Is there a preferred scale for the prior under M_2? The answer would seem to depend on *how the sample size is chosen*. If the sample size is merely a reflection of the available resources, it contains no information about the client's prior belief. In this

case a scale of the order of σ seems appropriate to us. This paper is based on such a tacit assumption.

However, the sample size may be chosen more judiciously to reflect the client's belief about what kind of deviations of $\bar{x}$ or μ from M_1 cannot be neglected. His sample size may then reflect his concern at the design stage that for this kind of alternative there is a reasonably high probability of choosing M_2. Indirectly, n now gives a lot of information on utility and prior. We feel priors with a scale of $\sigma/\sqrt{n}$ is more appropriate here than the intrinsic or Jeffreys's proper priors or the data dependent priors of Section 2.5. Which of the two scales is appropriate can only be determined by the client.

Many other choices of scale may be appropriate, depending on the context and available information.

5.3. *Alternative approaches*

Some people, e.g., Kadane (in discussion of Berger and Sellke, 1987) have argued in favour of an older procedure. One calculates a posterior credibility interval for μ under M_2 which is known to be robust with respect to choice of prior π given M_2. Then M_1 is rejected if and only if the interval does not contain the value stipulated by M_1, namely, zero. If π is the Lebesgue measure, this reduces to the usual frequentist test. We would refer to the reply of Berger and Sellke (1987) as to why this may not always be appropriate.

Alternative approaches have been proposed by Goutis and Robert (1998) and Bernardo (1999). The method of Goutis and Robert (1998) is, however, not fully automatic. We discuss in detail Bernardo's (1999) paper. Somewhat in the spirit of the method described in the previous paragraph, Bernardo (1999) assumes M_2 is true and chooses π as the reference prior, which happens to be the Lebesgue measure here. He then considers a decision problem with two decisions—accept M_1 or accept M_2, and introduces an interesting utility function based on the Kullback–Leibler numbers. This is a better utility or loss function than 0-1 loss *but does not help determine the scale in the sense of Section* 5.2, for one can consider the Kullback–Leibler distance between distributions of a single observation or n observations. Bernardo (1999) makes the second choice. It seems he is guided by "universal consensus" on Example 1 that a deviation of $\bar{x}$ of more than $2\sigma/\sqrt{n}$ or $3\sigma/\sqrt{n}$ should lead to rejection of M_1. No such consensus exists even among classical (frequentist) statisticians, leading to an asymptotic theory based on a $1/\sqrt{n}$ scale due to Pitman, Le Cam and others and a different asymptotic theory due to Bahadur and others, where the parameter space is kept fixed and one studies such things as the exponential rate at which error of first kind tends to zero (see Serfling, 1980, for references). Nor does such a consensus exist among Bayesians. For example, Jeffreys (1961) takes a point of view that leads to tests that are radically different from what Bernardo (1999) calls a consensus. This was the motivation of most of the work of Berger and Pericchi (1996a,b).

5.4. Sharp hypothesis

What happens if one replaces a sharp hypothesis M_1 by an interval J: $-\delta \leqslant \mu \leqslant \delta$, where δ is a small number? There is some discussion in Berger and Sellke (1987).

If we again take the view (as in Sections 5.1 and 5.2) that there is no information in n, and $\bar{x}$ is in the support of π and indicates some deviation from M_1, then $\bar{x}$ is usually outside J. The usual asymptotics for the IBFs and other related methods is now more difficult to justify because for $\bar{x}$ outside J, the likelihood under M_1 integrated over J is not well approximated by the likelihood at $\mu = 0$. The asymptotics remain valid if J is very small and n is large but not very large so that δ is $o(1/n)$. To see this let $m_1^J(\boldsymbol{x})$ be the marginal density of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ with respect to the uniform prior over J and $m_1(\boldsymbol{x})$ be the marginal density under sharp M_1 ($\mu = 0$). Then we have

$$\frac{m_1^J(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{1}{2\delta}\int_{-\delta}^{\delta} \exp\left[-\frac{n}{2}\mu^2 + n\bar{x}\mu\right] \mathrm{d}\mu$$

$$= \frac{1}{2\delta n}\int_{-\delta n}^{\delta n} \exp\left[-\frac{t^2}{2n} + \bar{x}t\right] \mathrm{d}t.$$

For moderate values of $\bar{x}$ this integral will be nearly one if δn is sufficiently small, i.e., if δ is $o(1/n)$.

A similar recommendation by Berger and Delampady (1987) that the replacement (of the sharp hypothesis by an interval J) may be made when $\delta\sqrt{n}$ is sufficiently small (say $< 1/4$) seems to be valid if $\sqrt{n}\bar{x}$ is moderate. Then the ratio above can be written as

$$\frac{m_1^J(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{1}{2\delta\sqrt{n}}\int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \exp\left[-\frac{t^2}{2} + (\sqrt{n}\bar{x})t\right] \mathrm{d}t.$$

5.5. A comment on methodology

Why not approach the problem of choosing a nonsubjective prior directly as Jeffreys does in his book for Examples 1 and 2? It is not easy to analyze other examples in this way. Moreover, even in Example 2 the Cauchy prior seems to be much more popular than the arguments of Jeffreys can justify. We have suggested if in a problem of this kind there is a prior which gives rise to a Bayes factor that is well approximated by an appealing data analytic heuristic procedure, then each of the two—the prior and the method—lends support to the other. In particular, we have shown that the Cauchy prior is obtainable in Example 2 in this way. These considerations are somewhat technical but how else can one see that the normal prior in Example 2, though so similar to the Cauchy prior, is quite inappropriate? It follows from Jeffreys's argument or directly that the Bayes factor with normal prior for μ and say the noninformative prior for σ does not go to infinity even if $\bar{x}$ tends to infinity.

Acknowledgements

Our understanding and, hopefully, the paper has improved because of critical comments of two anonymous referees and penetrating questions of an editor. We thank these three distinguished Bayesians.

The paper is our tribute to Professor C.R. Rao whose pioneering contributions include testing and model selection.

References

Atkinson, A.C., 1978. Posterior probabilities for choosing a regression model. Biometrika 65, 39–48.

Berger, J.O., 1997. P.C. Mahalanobis Memorial Lecture. Indian Statistical Institute, Calcutta.

Berger, J.O., Delampady, M., 1987. Testing precise hypothesis. Statistical Science 2, 317–352.

Berger, J.O., Pericchi, L.R., 1996a. The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. 91, 109–122.

Berger, J.O., Pericchi, L.R., 1996b. The intrinsic Bayes factor for linear models (with discussion). In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics, Vol. 5. Oxford University press, London, pp. 25–44.

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

Bernardo, J.M., 1980. A Bayesian analysis of classical hypothesis testing. In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics. Valencia: University Press, pp. 605–647.

Bernardo, J.M., 1999. Nested hypothesis testing: the Bayesian reference criterion. In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics. Vol. 6. Oxford University Press, London, pp. 101–130.

Bernardo, J.M., Bayarri, M.J., 1985. Bayesian model criticism. In: Florens, J.P. et al. (Eds.), Model Choice. Brussels: Pub. Fac. Univ. Saint Louis, pp. 43–59.

Box, G.E.P., Tiao, G.C., 1962. A further look at robustness via Bayes's theorem. Biometrika 49, 419.

Edwards, W., Lindman, H., Savage, L.J., 1963. Bayesian statistical inference for psychological research. Psychol. Rev. 70, 193–242.

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Amer. Statist. Assoc. 74, 153–160.

Goutis, C., Robert, C.P., 1998. Model choice in generalised linear models: a Bayesian approach via Kullback–Leibler projections. Biometrika 85, 29–37.

Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics, Vol. 4. Oxford University Press, London, pp. 147–167.

Ghosh, J.K., Bhanja, J., Purkayastha, S., Samanta, T., Sengupta, S., 1997. A statistical approach to geological mapping. Under revision. Mathematical Geology, to appear.

Ghosh, J.K., Ramamoorthi, R.V., 2000. Bayesian Nonparametrics. In preparation.

Heath, D., Sudderth, W., 1978. On finitely additive priors, coherence, and extended admissibility. Ann. Statist. 6, 333–345.

Iwaki, K., 1997. Posterior expected marginal likelihood for testing hypothesis. J. Econom. Asia Univ. 21, 105–134.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, London.

Lempers, F.B., 1971. Posterior Probabilities of Alternative Linear Models. University of Rotterdam Press, Rotterdam.

Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.

Morris, C., 1983. Parametric empirical Bayes Inference: Theory and applications. J. Amer. Statist. Assoc. 78, 47–65.

O'Hagan, A., 1995. Fractional Bayes factors for model comparisons. J. Roy. Statist. Soc. Ser. B 57, 99–138.

Pérez, J.M., 1998. Development of expected posterior prior distributions for model comparisons. Ph.D. Thesis, Purdue University, Lafayette, IN, submitted.

Regazzini, E., 1987. De Finetti's coherence and statistical inference. Ann. Statist. 15, 845–864.

San Martini, A., Spezzaferri, F., 1984. A predictive model selection criterion. J. Roy. Statist. Soc. Ser. B 46, 296–303.

Schervish, M.J., 1995. Theory of Statistics. Springer, New York.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Smith, A.F.M., Spiegelhalter, D.J., 1980. Bayes factors and choice criteria for linear models. J. Roy. Statist. Soc. Ser. B 42, 213–220.

Spiegelhalter, D.J., Smith, A.F.M., 1982. Bayes factors for linear and log-linear models with vague prior information. J. Roy. Statist. Soc. Ser. B 44, 377–387.