MODEL SELECTION WITH VAGUE PRIOR INFORMATION

(Asymptotics/Bayes factor/Hypothesis testing/Intrinsic priors/Model comparison)

ELÍAS MORENO*, F. JAVIER GIRON** AND M. LINA MARTÍNEZ**
** Departamento de Estadística e I.O. Facultad de Ciencias. Universidad de Málaga. Campus de Teatinos, s/n. 29071 Málaga. Spain.

ABSTRACT

In the Bayesian approach, the Bayes factor is the main tool for model selection and hypothesis testing. When prior information is weak, "default" or "automatic" priors, which are typically improper, are commonly used but, unfortunately, the Bayes factor is defined up to a multiplicative constant.

In this paper we revise some recent but already popular methodologies, intrinsic and fractional, to deal with improper priors in model selection and hypothesis testing. Special attention is paid to the intrinsic and fractional methods as tools devised to produce proper priors to compute actual Bayes factors.

Some illustration to hypothesis testing problems with more than one population are given, in particular the Behrens-Fisher problem is considered.

RESUMEN

En el enfoque bayesiano el factor de Bayes es la herramienta principal para la selección de modelos y para el contraste de hipótesis. Cuando la información a priori es escasa, generalmente se suelen utilizar distribuciones impropias, pero desafortunadamente, en este caso, el factor de Bayes queda definido salvo una constante multiplicative.

En este artículo se pasa revista a algunas de las recientes metodologías, a saber intrínseca y fraccional, para solventar el problema de la utilización de distribuciones a priori impropias en los problemas de selección de modelos y contraste de hipótesis, que sin embargo ya son habituales en la práctica. Se presta atención especial a los métodos intrínseco y fraccional, como herramientas diseñadas para proporcionar distribuciones a priori propias con las que calcular el factor de Bayes.

Lo anterior se ilustra con ejemplos de problemas de contraste de hipótesis, en particular con el problema de Behrens-Fisher.

1. Introduction

Suppose we observe a data vector \( x = (x_1, x_2, ..., x_n) \) and assume that they are i.i.d. replications from either model \( \mathcal{M}_1 : \{L_1(\theta), P(\theta)\} \) or \( \mathcal{M}_2 : \{L_2(\theta), P(\theta)\} \). The model selection problem is that of choosing between one of these two models.

For a given prior \( P \) on the set \( \{\mathcal{M}_1, \mathcal{M}_2\} \), and under the 0-1 loss function, it is easily seen that the optimal decision consists in choosing \( \mathcal{M}_1 \) if and only if the inequality

\[
P(\mathcal{M}_1|x) \geq 1
\]

holds. The ratio of the posterior probabilities of the models in (1) is given by

\[
\frac{P(\mathcal{M}_1|x)}{P(\mathcal{M}_2|x)} = \frac{m_1(x)}{m_2(x)} \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)},
\]

where \( m_i(x) = \int_{\theta} L_i(x|\theta) P(\theta) \, d\theta, \ i = 1, 2 \), is the marginal distribution of the data under \( \mathcal{M}_i \). Other interesting loss functions for model selection can be found in San Martini and Spezzaferri (1984) and Bernardo and Smith (1994).

The first factor appearing in the right hand side of (2), is known as the Bayes factor (BF) of model \( \mathcal{M}_1 \) against \( \mathcal{M}_2 \). It is usually denoted as \( B_{12}(x) \). It is the ratio of posterior to prior odds and it can be seen as a measure of evidence given by the data in favor of (or against) \( \mathcal{M}_1 \). The logarithm of \( B_{12}(x) \) is called by Good (1984) the weight of the evidence provided by the data.
The BF does not depend on the prior $P$ and this is seen as an attractive property by many authors (Box 1980, Kass and Raftery 1995, Sawa 1987).

For nested models, that is for models that satisfy

(i) $\Theta_1 \subset \Theta_2$,

(ii) $f_2(x|\Theta_2) = f_1(x|\Theta_1)$ if $\theta_2 = \theta_1$,

an important property of the BF is that it is consistent (O'Hagan 1994). That means that for any $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$,

$$\lim_{n \to \infty} \left[ P_{\theta_1} \right] B_{12}(x) = \infty, \quad \lim_{n \to \infty} \left[ P_{\theta_2} \right] B_{12}(x) = 0.$$ 

Consequently, for any prior $P$, it follows that asymptotically $P(M_1|x) = 1$ when sampling from $M_1$ and $P(M_2|x) = 1$ when sampling from $M_2$. Therefore, by using BF's we always choose the true model, asymptotically.

On the other hand, O'Hagan (1995) argued that the BF is highly sensitive to the prior distribution on the parameter in the model, even when the sample size increases. In fact, an approximation to the marginal density $m_1(x)$ is given by

$$m_1(x) = \pi(\hat{\theta}_1) f_1(\hat{\theta}_1) \left( 2\pi \right)^{-d_1/2} |V_1|^{1/2},$$

where $\hat{\theta}_1$ is the MLE of $\theta_1$ and $V_1$ is the modal dispersion matrix. This fact is particularly disturbing when the prior information on the parameters is weak, as assumed throughout this paper, and the prior $n_1(\theta_1)$ is replaced by a default prior $n^w_1(\theta_1)$. The difficulty arises because the default prior is typically improper, so that it must be written as

$$\pi^w_1(\theta_1) = c_1 h_1(\theta_1),$$

where $h_1(\theta_1)$ is a function whose integral diverges. Therefore, the normalizing constant $c_1$ cannot be specified and, as a consequence, the BF with improper priors is now written as

$$B_{12}^w(x) = \frac{c_1}{c_2} \frac{\int_{\theta_1} f_1(x|\theta_1) h_1(\theta_1) d\theta_1}{\int_{\theta_2} f_2(x|\theta_2) h_2(\theta_2) d\theta_2},$$

which is defined up to the multiplicative constant $c_1/c_2$.

To avoid such an indetermination is far from trivial. The Bayesian Information Criterion (BIC) proposed by Schwarz (1978), uses an asymptotic expansion of the marginal densities of the data, and the approximate Bayes factor is obtained by simply deleting the term that contains the priors. More precisely, the approximate Bayes factor $B_{12}^w(x)$ is given by

$$-2\log B_{12}^w(x) = -2\log \frac{f_1(x|\theta_1)}{f_1(x|\theta_2)} + (d_1 - d_2) \log n,$$

where $d_i$ is the dimension of the parameter space $\Theta_i$ and $\hat{\theta}_i(x)$ is the MLE of $\theta_i$. Notice that the BIC corrects the likelihood ratio by the factor $\exp\left(- (d_1 - d_2) \log n \right)$.


The simplest method is simply to set an initial condition on the BF to determine the arbitrary constant $c_1/c_2$. The idea was proposed by Smith and Spiegelhalter (1980) for nested model situations. They consider an «imaginary training sample» $x_0$ having the smallest sample size for which $0 < m_1(x_0) < \infty$, $i = 1, 2$ and such that it would most favor the simplest model. Then, $B_{12}^w(x_0)$ is set equal to one and so determines the constant $c_1/c_2$. This yields

$$c_1 = \int_{\theta_1} f_1(x_0|\theta_1) h_1(\theta_1) d\theta_1,$$

so that the resulting approximation, say $B_{12}^w(x)$, to the BF is

$$B_{12}^w(x) = B_{12}^w(x_0).$$

Unfortunately this procedure is biased toward the simpler model since the BF is set equal to one for a sample point that favors the simpler one.

Furthermore, even when the imaginary training sample is considered in a symmetric situation with respect to the two models, the resulting BF does not necessarily work well. Let us illustrate this point with the following simple example.

**Example 1.** Suppose we are interested in testing the hypotheses $H_1: X$ has a density $N(x \mid \theta, 1)$ with $\theta \leq 0$, versus $H_2: X$ has a density $N(x \mid \theta, 1)$ with $\theta \geq 0$. Assuming the conventional uniform prior for $\theta$, the models to be compared are,

$$M_1 : \{N(x \mid \theta_1, 1), \pi^w_1(\theta_1) = c_1 l_{[\theta_1 \leq \theta_0]}(\theta_1)\},$$

$$M_2 : \{N(x \mid \theta_2, 1), \pi^w_2(\theta_2) = c_2 l_{[\theta_2 \geq \theta_0]}(\theta_2)\}.$$ 

For a given data set $x = (x_1, x_2, ..., x_n)$, the BF turns out to be
where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \) and \( \Phi \) represents the cumulative standard normal distribution. An obvious condition on the BF would be that for \( \bar{x} = 0 \), both models are equally likely. This yields

\[
B_{12}^0(0, n) = 1,
\]

and therefore \( c_1 = c_2 \). The resulting BF is

\[
B_{12}(\bar{x}, n) = \frac{1 - \Phi(\bar{x} \sqrt{n})}{\Phi(\bar{x} \sqrt{n})}.
\]

Its asymptotic behavior is as follows. When sampling from \( p_\theta = N(\theta, 1) \), the distribution of \( \Phi(\bar{x} \sqrt{n}) \) is, for any \( 0 < u < 1 \), given by

\[
H_u(u) = P_\theta[\Phi(\bar{x} \sqrt{n}) \leq u] = P_\theta[\bar{x} \leq \Phi^{-1}(u)] = \Phi(\Phi^{-1}(u) - \theta \sqrt{n}).
\]

Thus,

\[
\lim_{n \to \infty} P_\theta[\Phi(\bar{x} \sqrt{n}) \leq u] = \begin{cases} 1 & \text{if } \theta < 0, \\ u & \text{if } \theta = 0, \\ 0 & \text{if } \theta > 0. \end{cases}
\]

Therefore,

\[
\lim_{n \to \infty} \left[ P_{\theta=0} \right] B_{12}(\bar{x}, n) = \frac{1 - U}{U},
\]

where \( U \) is a random variable uniform \((0, 1)\) distributed, and

\[
\lim_{n \to \infty} \left[ P_{\theta=0} \right] B_{12}(\bar{x}, n) = \begin{cases} 1 & \text{if } \theta < 0, \\ 0 & \text{if } \theta > 0. \end{cases}
\]

That is, consistency holds for values of \( \theta \) different from zero, but consistency does not hold when sampling from the \( N(0, 1) \) distribution.

In this example the likelihood ratio coincides with the Schwarz’s criterion and it is easy to see that

\[
\lim_{n \to \infty} \left[ P_{\theta=0} \right] B_{12}^\gamma(\bar{x}, n) = \begin{cases} 1 & \text{if } \theta < 0, \\ 0 & \text{if } \theta > 0, \end{cases}
\]

which shows a more reasonable asymptotic behavior than that of \( B_{12}(\bar{x}, n) \).

Furthermore, \( B_{12}(\bar{x}, n) \) as a measure of the evidence provided by the data in favor of \( M_1 \) is not acceptable. For
Bayes factors. Section 3 discusses intrinsic and fractional priors. These are priors for which the corresponding Bayes factors are, respectively, asymptotically equivalent to the intrinsic Bayes factor (IBF) and the fractional Bayes factor (FBF). Section 4 gives some applications of the intrinsic and fractional methodologies to some important testing problems related to more than one population. Section 5 gives some extensions and lists some open problems, and in Section 6 some concluding remarks are given.

2. Partial, Intrinsic and Fractional Bayes factors

The partial Bayes factor (PBF) was introduced by Leamer (1978) and the basic idea is as follows. The sample $x = (x_1, x_2, ..., x_n)$ is split into two parts $(x(l), x(n - l))$. Part $x(l)$, called training sample, is intended for converting the improper prior into a proper posterior, that is

$$
PI^n_i|\theta_0(x(l)) = \frac{f(x(l)|\theta_i) \pi^n_i(\theta)}{m^n_i(x(l))}, \quad i = 1, 2, \quad (3)
$$

where $x(l)$ is such that $0 < m^n_i(x(l)) < \infty$. With the remaining part of the data $x(n - l)$ the Bayes factor is computed by using (3) as the prior. This gives

$$
B_{12}^P(x) = \frac{\int f(x(n - l)|\theta_1) \pi^n_1(\theta_1|x(l))d\theta_1}{\int f(x(n - l)|\theta_2) \pi^n_2(\theta_2|x(l))d\theta_2}, \quad (4)
$$

which is called by O'Hagan (1995) a «partial Bayes factor». Equivalently, (4) can be written as

$$
B_{12}^P(x) = \frac{\int f(x|x_0)\pi^n_1(\theta_1) d\theta_1}{\int f(x|x_0)\pi^n_2(\theta_2) d\theta_2} \cdot m^n_1(x(l)) = B_{12}^P(x)B_{21}^P(x(l)). \quad (5)
$$

The PBF corrects the original Bayes factor $B_{12}^P(x)$ with the correction term $B_{21}^P(x(l))$, and by so doing the arbitrary constant $c_1/c_2$ cancels out in (5). It should be noted that for a given sample x we can typically consider different training samples $x(l)$. Therefore, there are a multiplicity of PBF, one for each training sample. The difficulty is to know which of them is to be chosen.

To avoid this difficulty Berger and Pericchi (1996b) first suggest considering all the possible training samples with minimal size. That is, all subsamples $x(l)$ such that $0 < m^n_i(x(l)) < \infty$ for which there is no proper subsample that satisfies these inequalities. They termed this training sample a minimal training sample. Second, they take the arithmetic mean of the PBF’s for all minimal training samples. This produces the so-called arithmetic intrinsic Bayes (IBF) factor, defined as

$$
B_{12}^{*A}(x) = B_{12}^P(x) \frac{1}{L} \sum_{l=1}^{L} B_{21}^P(x(l)), \quad (6)
$$

where $L$ is the number of minimal training samples contained in the sample. Other ways of «averaging» the PBF’s are possible. For instance, by taking a geometric mean, the geometric intrinsic Bayes factor results in

$$
B_{12}^{*G}(x) = B_{12}^P(x) \left( \prod_{l=1}^{L} B_{21}^P(x(l)) \right) \frac{1}{L}.
$$

By taking the median of the PBF’s, they obtain the «median» intrinsic Bayes factor. Notice that these empirical functions re-sample the sample and, although they are termed by their authors «Bayes factors», they are not actual Bayes factors. Their behavior is studied by Berger and Pericchi (1996a, 1996b, 1997a, 1997b) in various contexts.

An alternative way of avoiding the arbitrariness of choosing the training sample for which the PBF is computed, was suggested by O’Hagan (1995). Instead of using either a training sample, or all possible minimal training samples, he replaces the correction term $B_{21}^P(x(l))$ in (5) by

$$
F_{21}(x) = \frac{\int \int f(x|x_0) \pi^n_1(\theta_1) d\theta_1}{\int \int f(x|x_0) \pi^n_2(\theta_2) d\theta_2}. \quad (7)
$$

This way, he defines the fractional Bayes factor (FBF) as

$$
B_{12}^{*F}(x) = B_{12}^P(x) F_{21}(x).
$$

Other fractions than $\ln n$ can be considered. In fact, O’Hagan (1995) argues that a larger fraction would reduce sensitivity to the prior and he also proposes using the fractions long $n/h$ or $\sqrt{n}/n$. In favour of the fraction $\ln n$, however, many compelling arguments can be found in the literature (Berger and Mortera 1995, Kass and Wasserman 1995, Moreno 1997). We also remark that even the name of fractional Bayes factor is misleading as it is not really an actual Bayes factor.

Advantages and disadvantages of the above IBF’s and FBF’s have been extensively studied; see, in particular, the comparisons between the IBF and FBF by O’Hagan (1995, 1997), the criticisms on the stability of the IBF with respect to the sample by Bertolino and Racugno (1996), the review by De Santis and Spezzaferri (1997) and the criticisms and comparisons by Berger and Pericchi (1997a).

3. The intrinsic and fractional priors

From a theoretical viewpoint, the important question on the IBF and the FBF is to know whether it corresponds to an actual Bayes factor for sensible priors. If so, it provides a Bayesian justification for these methods, and consistency of these empirical Bayes factors is automatically satisfied.
With the so-called intrinsic priors, this question was solved asymptotically by Berger and Pericchi (1996b). These are priors $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ for which the corresponding Bayes factor.

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

and $B_{21}(x)$ are asymptotically equivalent under the two models $M_1$ and $M_2$.

By equating the limit of $B_{21}(x)$ and $B_{21}(x)$ as $n \to \infty$ under the two models

$$B_{21}(x) = B_{21}^M(x) \left[1 + O_p(1)\right], \quad i = 1, 2;$$

Berger and Pericchi (1996b) showed that the intrinsic priors satisfy the functional equations

$$E_{\pi_1}^N B_{21}^N(x(l)) = \frac{\pi_1(\psi_2(\theta_1))}{\pi_1(\psi_2(\theta_1))} \frac{\pi_1^N(\theta_1)}{\pi_1(\theta_1)}, \quad (8)$$

$$E_{\pi_2}^N B_{21}^N(x(l)) = \frac{\pi_2(\theta_2)}{\pi_2(\theta_2)} \frac{\pi_2^N(\psi_1(\theta_2))}{\pi_1(\psi_1(\theta_2))}. \quad (9)$$

The expectations in equations (8) and (9) are taken with respect to $f_l(x(l)|\theta_1)$ and $f_2(x(l)|\theta_2)$, respectively; $\psi_2(\theta_1)$ denotes the limit of the maximum likelihood estimator $\hat{\theta}_2(x)$ under model $M_1$ at the point $\theta_1$, and $\psi_1(\theta_2)$ the limit of $\hat{\theta}_1(x)$ under model $M_2$ at the point $\theta_2$ (see also Cox (1961), Huber (1967) and Dmochowsky (1994)).

Unfortunately, these equations may have many solutions or may not have a solution.

When the models under comparison are nested, say $f_1(x(0)|\theta_1)$ is nested in $f_2(x|\theta_2)$, the intrinsic priors always exist, under some mild conditions, and they are shown to be any pair (Moreno, Bertolino and Racugno 1998b), $(\pi_1(\theta_1), \pi_2(\theta_2))$ such that

(a) $\pi_1(\theta_1)$ is any prior in the class

$$\Gamma_1 = \left\{ \pi_1(\theta_1) : \int_{\theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\theta_1} T(\theta_1) \pi_1(\psi_1(\theta_1)) d\theta_2 = 1 \right\}$$

where $T(\theta_1) = \pi_1^N(\psi_2(\theta_1)) / \pi_1^N(\psi_1(\theta_1)) E_{\pi_1}^N B_{21}^N(x(l))$.

(b) For each $\pi_2(\theta_2) \in \Gamma_1$, $\pi_2(\theta_2)$ is given by

$$\pi_2(\theta_2) = T(\theta_2) \pi_2(\psi_1(\theta_2)).$$

Similarly, if the nested models satisfy that

$$F_{12}(\theta_2) = \lim_{n \to \infty} \left[ \frac{\int_{\theta_2} f_1(x|\theta_2)^{1/n} \pi_1^N(\theta_2) d\theta_2}{\int_{\theta_2} f_2(x|\theta_2)^{1/n} \pi_2^N(\theta_2) d\theta_2} \right],$$

is a degenerate random variable, the fractional priors $(\pi_1(\theta_1), \pi_2(\theta_2))$—priors for which their associated Bayes factor and $B_{21}^N(x)$ are asymptotically equivalent under the two models $M_1$ and $M_2$—are the solutions to the functional equation

$$F_{12}^N(\theta_2) = \frac{\pi_2^N(\theta_2)}{\pi_2^N(\psi_1(\theta_2))} \frac{\pi_1^N(\psi_1(\theta_2))}{\pi_1^N(\theta_2)}. \quad (10)$$

It can be shown (Moreno 1997) that the solutions to this equation are any pair $(\pi_1(\theta_1), \pi_2(\theta_2))$, where $\pi_1(\theta_1)$ is any member of the class

$$\Gamma_2 = \left\{ \pi_1(\theta_1) : \int_{\theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\theta_1} S(\theta_1) \pi_1(\psi_1(\theta_1)) d\theta_1 = 1 \right\},$$

with

$$S(\theta_1) = F_{12}^N(\theta_2) \frac{\pi_1^N(\theta_2)}{\pi_1^N(\psi_1(\theta_2))},$$

and for each $\pi_1(\theta_1) \in \Gamma_2$,

$$\pi_2(\theta_2) = S(\theta_1) \pi_1(\psi_1(\theta_1)).$$

Unfortunately, the intrinsic and fractional priors are classes of distributions for which robustness of their Bayes factor cannot be asserted (Moreno 1997, Moreno Bertolino and Racugno 1998).

A procedure for choosing intrinsic and fractional priors in their respective classes, was proposed by Moreno, Bertolino and Racugno (1998b). Their procedure considers the restriction of the improper prior for the simplest model, say $\pi_1^N(\theta_1)$, to a subset on which it integrates to a finite quantity. More precisely, the proper prior $\pi_1^N(\theta_1)$ is $\pi_1^N(\theta_1)/k_n 1_{C_n}(\theta_1)$ considered, where $C_n$ is an increasing sequence of subsets of the parameter space $\Theta_1$ such that $\int_{\theta_1} \pi_1^N(\theta_1) d\theta_1 = k_n$ and $\lim_{n \to \infty} C_n = \Theta_1$. For this prior the corresponding $\pi_1^N(\theta_2)$ is then computed. Then, $\pi_2^N(\theta_2) = \lim_{n \to \infty} \pi_2^N(\theta_2)$. It can be shown that the latter prior is a unique density. This procedure gives the intrinsic prior

$$\left\{ \pi_1^N(\theta_1), \pi_2^N(\theta_2) E_{\pi_1}^N B_{21}^N(x(l)) \right\}.$$

The same procedure applies with the fractional methodology but, unfortunately there not always exists a probability density $\pi_1(\theta_1)$. Whenever it exists, the fractional priors are given by

$$\left\{ \pi_1^N(\theta_1), \pi_2^N(\theta_2) F_{12}(\theta_2) \right\}.$$
The resulting Bayes factor for the intrinsic priors is

\[
B_{21}^I(x) = \int_{0}^{\infty} f_2(x|\theta_2) \pi_2^N(\theta_2) E_{\theta_2}^{M_2} B_{21}^I(x|l) d\theta_2,
\]

and, for the fractional priors,

\[
B_{21}^F(x) = \int_{0}^{\infty} f_2(x|\theta_2) \pi_2^N(\theta_2) F_{\theta_2}(\theta_2) d\theta_2.
\]

Let us illustrate the procedure with a simple example.

**Example 2.** Suppose we want to compare model

\[ M_1 : N(x|\mu_1, \sigma_1^2) \]

versus

\[ M_2 : N(x|\mu_2, \sigma_2^2) \]

Here the prior for model \( M_1 \) is concentrated on \( \theta = 0 \), and the prior for model \( M_2 \) is an improper uniform distribution.

It is easy to see that the intrinsic priors are

\[ \pi_2^N(\theta) = 1_{\theta=0}(\theta) \]

However, the fractional priors do not exist. Indeed, the resulting fractional priors are

\[ \pi_2^F(\theta) = 1_{\theta=0}(\theta) \]

But

\[ \int_{0}^{\infty} \pi_2^F(\theta) d\theta < 1, \]

does not integrate up to one. Therefore, in this example, the fractional prior for model \( M_2 \) is not a probability density.

4. AN APPLICATION TO THE BEHRENS-FISHER PROBLEM

Testing problems for which frequentist methods fail, can be solved with the help of the intrinsic and fractional methodologies. A famous example is the well-known Behrens-Fisher problem.

Let \( N(x|\mu_1, \sigma_1^2) \), \( N(x|\mu_2, \sigma_2^2) \) be two normal distributions where the means \( \mu_1, \mu_2 \) and variances \( \sigma_1^2, \sigma_2^2 \) are unknown. Samples \( x = (x_1, x_2, ..., x_n) \), \( y = (y_1, y_2, ..., y_n) \), of sizes \( n_1 \) and \( n_2 \), respectively, are taken, and the samples means and variances are denoted by \( \bar{x} \) and \( \sigma_1^2 \) and \( \bar{y} \) and \( \sigma_2^2 \), respectively.

The Behrens-Fisher problem consists in testing \( H_0 : \mu_1 = \mu_2 \) against one of the alternatives \( \mu_1 < \mu_2, \mu_1 > \mu_2 \) or \( \mu_1 \neq \mu_2 \), assuming that \( \sigma_1^2, \sigma_2^2 \) are unrelated. Here we focus on the alternative \( H_1 : \mu_1 > \mu_2 \).

Under the frequentist point of view this problem poses the difficulty that the normal-theory linear model cannot be applied because of the presence of the two unrelated variances \( \sigma_1^2, \sigma_2^2 \). Some approximations to the solution have been given by Fisher (1936), Wald (1955), and Welch (1947), and their relative merits are discussed in Mehta and Srinivasan (1970) and Pfanzagl (1974).

Under the Bayesian viewpoint the Behrens-Fisher problem can be focused as one of model comparison. The models to be compared are

\[ M_1 : f_1(x|\theta_1) = N(x|\mu_1, \sigma_1^2) \]

\[ M_2 : f_2(x|\theta_2) = N(x|\mu_2, \sigma_2^2) \]

where \( z = (x, y) \), \( \theta_1 = (\mu_1, \tau_1, \tau_2) \) and \( \theta_2 = (\mu_2, \sigma_1^2, \sigma_2^2) \). The usual priors for \( \mu_1, \mu_2, \sigma_1, \log \sigma_2 \) are independent uniform distributions over \((-\infty, \infty)\). (Jeffreys, 1961). They are appropriated for inference purposes on the \( \mu_1 \) and, in fact, the posterior distribution of \( \mu_1 - \mu_2 \) is a linear transformation of the so-called Behrens-Fisher distribution (Lindley, 1970). However, these priors are improper and can be written as

\[ \pi_1^N(\theta_1) = \frac{c_1}{\tau_1 \tau_2} 1_{\mathbb{R}^{+}}(\mu_1, \tau_1, \tau_2), \]

\[ \pi_2^N(\theta_2) = \frac{c_2}{\sigma_1 \sigma_2} 1_{\mathbb{R}^{+}}(\mu_1, \mu_2, \sigma_1, \sigma_2), \]

thus depending on arbitrary positive constants \( c_1, c_2 \), which make them unsuitable for testing \( H_0 \) versus \( H_1 \). However, intrinsic priors can be derived for the Behrens-Fisher problem.

**Lemma 1.** (Moreno, Bertolino and Racugno, 1997)

For each point \( \mu, \tau_1, \tau_2 \) the intrinsic prior of \( \theta_2 \) is

\[ \pi_2^I(\theta_2|\mu, \tau_1, \tau_2) = \frac{1}{\mathbb{R}^{n_1+n_2}} \prod_{i=1}^{n_1} N \left( \mu, \frac{\tau_1^2 + \sigma_1^2}{2} \right) \]

where \( HC^*(\sigma|0, \tau) \) denotes the half Cauchy density.

Integrating out the variables \( (\mu, \tau_1, \tau_2) \) with respect to the default prior \( \pi_2^N(\theta_1) \), renders the unconditional intrinsic prior of \( \theta_2 \),

\[ \pi_2^I(\theta_2) = \int_{\mathbb{R}^{n_1+n_2}} \frac{1}{\mathbb{R}^{n_1+n_2}} \prod_{i=1}^{n_1} N \left( \mu, \frac{\tau_1^2 + \sigma_1^2}{2} \right) HC^*(\sigma|0, \tau) \]

\[ \frac{1}{\tau_1 \tau_2} \]
Similarly, the fractional priors can also be obtained.

Lemma 2. (Moreno, Bertolino and Racugno, 1997)

For each point \( \mu, \tau_1, \tau_2 \), the fractional prior of \( \theta_2 \) is given by

\[
\pi_i^F(\theta_2 | \mu, \tau_1, \tau_2) = \prod_{i=1}^{2} N\left( \mu, \mu, \frac{\tau_i^2}{2} \right) HN^+\left( \sigma, 0, \frac{\tau_i^2}{2} \right)
\]

where \( HN^+\left( \sigma, 0, \frac{\tau_i^2}{2} \right) \) denotes a half Normal density.

The intrinsic \( \pi_i^I(\mu, \mu, \sigma_1, \sigma_2, \mu, \tau_1, \tau_2) \) and fractional \( \pi_i^F(\mu, \mu, \sigma_1, \sigma_2, \mu, \tau_1, \tau_2) \) priors are quite reasonable. The parameters \( (\mu_i, \sigma_i) \), \( i = 1, 2 \), are exchangeable in both the intrinsic and the fractional priors. This is sensible as long as the labels given to the populations are somewhat arbitrary.

The main differences between the intrinsic and fractional priors \( \pi_i^I(\theta_2 | \mu, \tau_1, \tau_2) \) and \( \pi_i^F(\theta_2 | \mu, \tau_1, \tau_2) \) are that the tails of the densities of the \( \sigma_i \)'s are thicker for the intrinsic than for the fractional, and that \( (\mu_i, \sigma_i) \) are dependent in the intrinsic whereas they are independent in the fractional prior.

With the above intrinsic and fractional priors the corresponding Bayes factors \( B_{21}^I(x, y) \), and \( B_{21}^F(x, y) \), can be computed.

For comparison purposes we also consider the Schwarz approximation. It can be shown (Moreno, Bertolino and Racugno 1997), that

\[
B_{21}^F(x, y) = \frac{1}{n} \left[ 1 + \frac{(x - \hat{\mu})^2}{s_x^2} \right] \left[ 1 + \frac{(\bar{y} - \hat{\mu})^2}{s_y^2} \right] \exp\left\{ - \frac{1}{2} \log \frac{n_1 n_2}{n} \right\}
\]

where \( n = n_1 + n_2 \) and \( \hat{\mu} \) is the solution to the likelihood equation

\[
\sum_{i=1}^{z} \frac{n_i^2 (\bar{x}_i - \hat{\mu})^2}{s_i^2 + n_i (\bar{x}_i - \hat{\mu})^2} = \hat{\mu} \sum_{i=1}^{z} \frac{n_i^2 (\bar{x}_i - \hat{\mu})^2}{s_i^2 + n_i (\bar{x}_i - \hat{\mu})^2}
\]

that is, the point that maximizes the likelihood function of model \( M_1 \).

Let us illustrate the behavior of the above Bayes factors with the data taken from Box and Tiao (1973, example 2.5.A., p. 107).\n
\[
\begin{align*}
\text{n}_1 &= 20, \quad s_x^2 = 12, \\
\text{n}_2 &= 12, \quad s_y^2 = 40,
\end{align*}
\]

and several values of \( \bar{x} - \bar{y} \). In Table 2, the corresponding values of \( B_{21}^I, B_{21}^F, \) and \( B_{21}^t \) are displayed.

<table>
<thead>
<tr>
<th>( \bar{x} - \bar{y} )</th>
<th>P-values (Welch)</th>
<th>( B_{21}^I )</th>
<th>( B_{21}^F )</th>
<th>( B_{21}^t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.36</td>
<td>0.15</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>2.20</td>
<td>0.69</td>
<td>0.29</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>4.22</td>
<td>3.10</td>
<td>1.04</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>6.33</td>
<td>1.96</td>
<td>2.49</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. P-values of \( B_{21}^I, B_{21}^F, \) and \( B_{21}^t \)

Table 2 shows that \( B_{21}^F \) and \( B_{21}^t \) have a similar behavior. For small values of the difference \( \bar{x} = \bar{y} \), they clearly favour \( M_1 \) and as the difference grows the model \( M_2 \) is strongly favored. That behavior is sensible.

The Schwarz approximation \( B_{21}^F \) favours the complex model \( M_2 \) for small departures of \( \bar{x} = \bar{y} = 5 \) from zero. In particular, for \( \bar{x} = \bar{y} = 5 \) the evidence in favour of \( M_2 \) is larger than the one given by \( B_{21}^F \) and \( B_{21}^t \).

For \( \alpha = 0.05 \), the Welch approximation gives the critical region \( W_{0.05} = \{ x, y : |x - y| \geq 4.22 \} \). This conclusion is in agreement with the Schwarz approach but is in disagreement with those given by either \( B_{21}^I \) or \( B_{21}^F \).

5. EXTENSIONS AND OPEN PROBLEMS

The application of the intrinsic and fractional methodologies to non-regular models have been discussed by several authors. The kind of non-regular models considered include models with improper likelihoods as those appearing in mixture models, likelihoods where the sample space depends on the parameter space, models that have an increasing multiplicity of parameters as in the Neyman-Scott problem, or non-nested models as that considered in Example 1. We remark, however, that non-nested models have not yet an acceptable solution nowadays. Some particular cases, for instance, the class of models that have a given invariance structure are an exception (Berger, Pericchi and Varshavsky (1996) and Moreno, Bertolino and Racugno (1998b)).

Some solutions to the problems listed above appear in the papers by Berger and Pericchi (1997), Shui (1998), Moreno and Liseo (1998) and Moreno, Bertolino and Racugno (1999), and we do not review them here.

Another exception is that of testing nonnested models for which there exist a model nested to both. Example 1 is in this situation. In this case, the problem can be solved via
transforming it into an equivalent testing problem for a nested model. Let us illustrate this point.

Example 1 (continued). By considering the model $M_3: \{X \text{ has density } N(\mu, 1)\}$, which is nested to $M_1$ and also to $M_2$, we can consider the Bayes factor

$$B_{12}(\bar{x}_n, n) = B_{13}(\bar{x}_n, n) B_{32}(\bar{x}_n, n),$$

where we are assuming that the prior $\pi_3$ appearing in the Bayes factor $B_{13}(\bar{x}_n, n)$, is exactly the same than that appearing in the Bayes factor $B_{32}(\bar{x}_n, n)$.

In Example 2 the intrinsic priors for comparing model $M_3$ versus $M_1$ was seen to be

$$\pi_1'(\theta) = \pi_1(\theta) \left( \frac{\int N(\theta, 1)}{\int N(\theta, 1)} \right),$$

where

$$\pi_1(\theta) = \frac{\int N(\theta, 1)}{\int N(\theta, 1)} \int \pi_0(\theta).$$

It is easy to see that the intrinsic priors for comparing model $M_3$ versus $M_2$ are given by

$$\pi_2'(\theta) = \pi_2(\theta) \left( \frac{\int N(\theta, 1)}{\int N(\theta, 1)} \right),$$

where

$$\pi_2(\theta) = \frac{\int N(\theta, 1)}{\int N(\theta, 1)} \int \pi_0(\theta).$$

Therefore, for a given sample $x = (x_1, x_2, \ldots, x_n)$, the Bayes factor for intrinsic priors $B_{12}'(\bar{x}_n, n)$, turns out to be

$$B_{12}'(\bar{x}_n, n) = \int \frac{a(x, \bar{x}_n, n)}{\Phi(-x_n + n \bar{x})} \Phi(-x) \text{d}x,$$

where

$$a(x, \bar{x}_n, n) = \exp\left( -x^2 \left( \frac{n}{2(1+n)} + \frac{1}{2} \right) + x \frac{n \bar{x}}{1+n} \right).$$

The asymptotic behavior of $B_{12}'(\bar{x}_n, n)$ is

$$\lim_{n \to \infty} [p_{\theta}] B_{12}'(\bar{x}_n, n) = \begin{cases} \infty \text{ if } \theta < 0, \\ 1 \text{ if } \theta = 0, \\ 0 \text{ if } \theta > 0, \end{cases}$$

so that the Bayes factor for the nonnested models $M_1$ and $M_2$, with the intrinsic priors is consistent.

For some values of $(\bar{x}_n, n)$ the resulting values of $B_{12}'(\bar{x}_n, n)$ are given in the second column of Table 3. For the prior $P(M_1) = P(M_2) = 1/2$, the posterior probabilities of model $M_1$, say $P^I(M_1|\bar{x}_n, n)$, are given in the third column of Table 3. The Schwarz's criterion to the Bayes factor and their corresponding posterior probabilities of model $M_1$ are displayed in the fourth and fifth columns, respectively.

| $(\bar{x}_n, n)$ | $B_{12}'(\bar{x}_n, n)$ | $P^I(M_1|\bar{x}_n, n)$ | $P^S(M_1|\bar{x}_n, n)$ |
|------------------|------------------------|------------------------|------------------------|
| $(-3, 1)$        | 7.1                    | 0.88                   | 90.0                   |
| $(-2, 1)$        | 3.2                    | 0.6                    | 7.4                    |
| $(-1, 1)$        | 1.7                    | 0.63                   | 1.6                    |
| $(0, 1)$         | 1.0                    | 0.50                   | 1.0                    |

Table 3. Values of $B_{12}'(\bar{x}_n, n)$

For values of the sample mean less or equal than three standard deviation from the mean $\theta$ of the alternative model, the intrinsic posterior probabilities of the model $M_1$ are more reasonable than those given by the Schwarcz's approach. For instance, for $(\bar{x}_n, n) = (-3, 1)$, the posterior probability with the intrinsic prior is 0.88 while with the Schwarcz it is 0.99, which seems to be too large.

We recall that for this example there are no fractional priors.

6. CONCLUSIONS

Improper priors are useful tools when prior information is weak. Furthermore, they are seen by many authors as the basic tools for doing «objective» or «scientific» inference. Our position is that a better analysis of a problem is obtained by incorporating prior information as described by a probability density. However, we recognize that a carefully elicitation of the prior distributions is not always possible and then automatic methods are necessary.

However, improper priors are not suitable for a direct application of the Bayes theorem to compute posterior probabilities. The prize to pay is a more elaborated analysis to derive a proper prior from the improper with which we start.

This is, in fact, what the training sample device suggests. We should say that even when empirical measures of evidence, as the variety of partial Bayes factor presented in this paper, are interesting in their own, we feel that we are more in agreement with the Bayesian paradigm if we use these partial Bayes factors as a tool to derive sensible prior distributions. In other words, we should act in agreement with the Berger and Pericchi's principle stated in the Introduction.

In practice, Bayes factors with fractional and intrinsic priors are very close to each other. While it seems that fractional priors are easier to be obtained than the intrinsic priors (because to calculate a limit in probability is typically less involved than an expectation), they may not always exist and this alone is a strong reason in favor of the intrinsic priors.

Interesting enough is the fact that many challenged problems, for which the standard classical theory does not
apply, do have, however, a default and sensible Bayesian solution.

ACKNOWLEDGMENTS

This research was supported by grants PB96-1388 and PB97-1403-C03 from the Ministry of Education, Spain.

REFERENCES


46. Welch, B.L. (1947) «The generalization of Student's problem when several populations are involved», Biometrika, 34, 28-35.