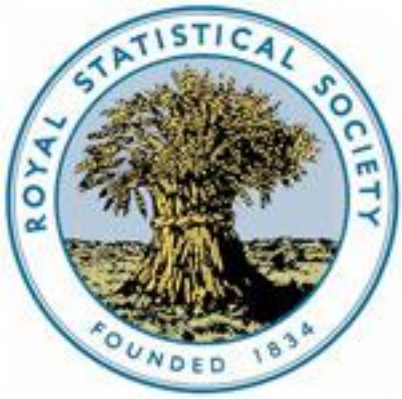


WILEY

Fractional Bayes Factors for Model Comparison
Author(s): Anthony O'Hagan
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1 (1995), pp. 99-138
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2346088
Accessed: 29-10-2016 05:59 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Royal Statistical Society, *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series B (Methodological)*

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 99–138

Fractional Bayes Factors for Model Comparison

By ANTHONY O'HAGAN†

University of Nottingham, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 16th, 1994,* Professor V. S. Isham *in the Chair*]

SUMMARY

Bayesian comparison of models is achieved simply by calculation of posterior probabilities of the models themselves. However, there are difficulties with this approach when prior information about the parameters of the various models is weak. Partial Bayes factors offer a resolution of the problem by setting aside part of the data as a training sample. The training sample is used to obtain an initial informative posterior distribution of the parameters in each model. Model comparison is then based on a Bayes factor calculated from the remaining data. Properties of partial Bayes factors are discussed, particularly in the context of weak prior information, and they are found to have advantages over other proposed methods of model comparison. A new variant of the partial Bayes factor, the fractional Bayes factor, is advocated on grounds of consistency, simplicity, robustness and coherence.

Keywords: ASYMPTOTIC NORMALITY; BAYESIAN INFERENCE; CONSISTENCY; FRACTIONAL BAYES FACTOR; MODEL CRITICISM; PARTIAL BAYES FACTOR; ROBUSTNESS

1. INTRODUCTION

1.1. *Bayes Factors*

Two models are proposed for data $\mathbf{x}$. Under model M_i, the data are related to parameters $\boldsymbol{\theta}_i$ by a distribution $f_i(\mathbf{x}\,|\,\boldsymbol{\theta}_i)$, and the prior distribution is $\pi_i(\boldsymbol{\theta}_i)$, $i=1, 2$. The posterior odds in favour of model 1 against model 2 can be written

$$\frac{P(M_1\,|\,\mathbf{x})}{P(M_2\,|\,\mathbf{x})}=\frac{P(M_1)}{P(M_2)}\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}=\frac{P(M_1)}{P(M_2)}B(\mathbf{x})$$

where $B(\mathbf{x})$ is known as the Bayes factor (in favour of model 1 against model 2) and

$$q_i(\mathbf{x})=\int \pi_i(\boldsymbol{\theta}_i)\,f_i(\mathbf{x}\,|\,\boldsymbol{\theta}_i)\,\mathrm{d}\boldsymbol{\theta}_i$$

is the marginal density of $\mathbf{x}$ under model i. The posterior odds are therefore the prior odds multiplied by the Bayes factor, and the Bayes factor can be seen as representing the weight of evidence in the data in favour of model 1 against model 2. If model 1 fits the data better than model 2, in the sense that $q_1(\mathbf{x})>q_2(\mathbf{x})$, then $B(\mathbf{x})>1$ and the posterior odds in favour of model 1 will be greater than the prior odds.

When prior information is weak, however, there are difficulties with the use of Bayes factors. The most obvious problem arises if we try to represent prior ignorance

†*Address for correspondence*: Department of Mathematics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK.
E-mail: aoh@uk.ac.nott.maths

© 1995 Royal Statistical Society 0035-9246/95/57099

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

about θ_1 and/or θ_2 by using improper priors. An improper prior for θ_i is usually written as

$$\pi_i(\theta_i) \propto h_i(\theta_i)$$

where h_i is a function whose integral over the θ_i-space diverges. For instance, a uniform prior would be expressed as $\pi_i(\theta_i) \propto 1$. Formally, we can write

$$\pi_i(\theta_i) = c_i h_i(\theta_i), \tag{1}$$

although the normalizing constant c_i does not exist, but treating it as an unspecified constant. This approach is common in Bayesian analysis of a single model, since the posterior distribution of the parameter θ_i is

$$\pi_i(\theta_i | \mathbf{x}) = \frac{\pi_i(\theta_i) f_i(\mathbf{x} | \theta_i)}{q_i(\mathbf{x})} = \frac{h_i(\theta_i) f_i(\mathbf{x} | \theta_i)}{\int h_i(t) f_i(\mathbf{x} | t) \, \mathrm{d}t}, \tag{2}$$

the constant c_i cancelling out. Provided that the integral in the denominator converges, the posterior density is well defined despite c_i being unspecified. However, if the prior distribution for model 1 is given in the improper form (1) the Bayes factor is

$$B(\mathbf{x}) = \frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} = c_1 \frac{\int h_1(\theta_1) f_1(\mathbf{x} | \theta_1) \, \mathrm{d}\theta_1}{\int \pi_2(\theta_2) f_2(\mathbf{x} | \theta_2) \, \mathrm{d}\theta_2}$$

and the unspecified c_1 does not cancel out. The Bayes factor is directly proportional to c_1. If improper prior distributions are given to both models,

$$B(\mathbf{x}) = \frac{c_1}{c_2} \frac{\int h_1(\theta_1) f_1(\mathbf{x} | \theta_1) \, \mathrm{d}\theta_1}{\int h_2(\theta_2) f_2(\mathbf{x} | \theta_2) \, \mathrm{d}\theta_2} \tag{3}$$

and so depends on the ratio c_1/c_2 of two 'unspecified constants'.

Various approaches have been advocated for dealing with this problem. One is simply to reject improper prior distributions, insisting that model comparison (and perhaps inference generally) is not meaningful unless genuine prior information is represented by proper prior distributions. Although this strict line will indeed remove the specific indeterminacy described above, I shall argue in Section 5 that the problem is more deep seated, and that Bayes factors are inherently sensitive to errors of specification of prior distributions.

1.2. *Imaginary Minimal Experiment*

A second approach is to remove the indeterminacy by a kind of thought experiment, as proposed by Spiegelhalter and Smith (1982). The basic idea is that, if we can imagine a specific set of data $\mathbf{x}_0$ such that we are prepared to assign a particular value to $B(\mathbf{x}_0)$, then c_1 or c_1/c_2 is thereby determined. If both prior distributions are improper, equation (3) becomes

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$B(\mathbf{x}) = B(\mathbf{x}_0) \frac{\int h_2(\theta_2) f_2(\mathbf{x}_0 | \theta_2) \, d\theta_2}{\int h_1(\theta_1) f_1(\mathbf{x}_0 | \theta_1) \, d\theta_1} \frac{\int h_1(\theta_1) f_1(\mathbf{x} | \theta_1) \, d\theta_1}{\int h_2(\theta_2) f_2(\mathbf{x} | \theta_2) \, d\theta_2}.$$

Spiegelhalter and Smith developed this argument in the context of nested linear models by supposing first that $\mathbf{x}_0$ derives from a *minimal* experiment. The concept of a minimal experiment is not precisely defined, but it should have the smallest number of observations consistent with obtaining proper posterior distributions $\pi_i(\theta_i | \mathbf{x}_0)$ from equation (2) for both $i=1$ and $i=2$. Their second step is to choose $\mathbf{x}_0$ to give maximal support to the simpler model, i.e. if model 1 is the simpler model $\mathbf{x}_0$ is chosen to maximize $B(\mathbf{x}_0)$. (This can of course be done independently of the unspecified c_1 or c_1/c_2.) Finally, they argued that if the experiment is minimal then intuitively it can at best give only very slight support to the simpler model, and thereby justify setting $B(\mathbf{x}_0)=1$ (as at least an approximation).

There are several difficulties with this method, not the least of which is the justification of setting $B(\mathbf{x}_0)=1$, but I shall concentrate here on the notion of a minimal data set. Consider a very simple regression situation in which model 2 asserts that observations $x_1, x_2, \ldots, x_n$ are independent given an unknown regression parameter β, with distributions $x_j | \beta \sim N(\beta a_j, 1)$. The a_js are values of the regressor variable. Model 1 asserts further that $\beta=0$ so that the x_js become independent standard normal observations. Under model 2, a uniform prior distribution $h_2(\beta)=1$ is proposed, with an unspecified proportionality constant c_2. Clearly, a minimal experiment needs one observation x_0 and a corresponding value $a_0 \neq 0$ of the regressor variable. However, it is not clear what value of a_0 would make this experiment 'minimal'. The guidance given by Spiegelhalter and Smith (1982), in a similar regression context, is to advocate that $|a_0|$ be as large as possible, but this would seem to make the experiment *maximally* informative about β (subject to being of minimal size).

To pursue this point we can obtain $B(x_0)$. The numerator $q_1(x_0)$ is the marginal density of x_0 under model 1, which is simply $(2\pi)^{-1/2} \exp(-x_0^2/2)$. The denominator is

$$q_2(x_0) = c_2 \int (2\pi)^{-1/2} \exp\left\{ -\frac{(x_0 - \beta a_0)^2}{2} \right\} d\beta = \frac{c_2}{|a_0|}.$$

Therefore

$$B(x_0) = c_2^{-1} |a_0| (2\pi)^{-1/2} \exp(-x_0^2/2).$$

This is maximized for any a_0 by $x_0=0$. Setting the resulting Bayes factor to 1 and $|a_0|$ to ∞, since there is no limit to how large a_0 can be in the imaginary experiment (even if that is not true in reality), gives $c_2=\infty$ and therefore $B(\mathbf{x})=0$ whatever the actual observed data $\mathbf{x}$ may be. The advice here seems to lead to a posterior certainty that model 2 is true, regardless of the actual data. In fact Spiegelhalter and Smith (1982) arbitrarily constrained $|a_0| \leqslant 1$ and so obtained a finite c_2, although there does not seem to be any justification for this constraint.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

A more literal idea of a minimal experiment would surely set $|a_0|$ as *small* as possible, to be minimally informative. This is no better, for then $c_2 = 0$ and $B(\mathbf{x}) = \infty$, leading to posterior certainty that model 1 is true, regardless of the actual data.

The point of this example is that, except in special circumstances, there is great ambiguity over the definition of a minimal experiment. The method simply does not resolve the indeterminacy of Bayes factors with improper priors.

1.3. *Asymptotics*

Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and the x_js are independent and identically distributed given $\boldsymbol{\theta}_i$ under model i with common density g_i, i.e.

$$f_i(\mathbf{x}|\boldsymbol{\theta}_i) = \prod_{j=1}^{n} g_i(x_j|\boldsymbol{\theta}_i).$$

Expanding now around the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_i$,

$$\log f_i(\mathbf{x}|\boldsymbol{\theta}_i) = \log L_i - (n/2)(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)' V_i^{-1}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i) + R,$$

where $L_i = f_i(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)$ is the maximized likelihood, $-nV_i^{-1}$ is the Hessian matrix of $\log f_i$ at $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ and R is the remainder term. For $|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i|$ of order $n^{-1/2}$, the remainder term is of order $n^{-1/2}$ and can be ignored for large n (and for larger $|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i|$ the likelihood $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ is itself negligible). To the same order of accuracy, the prior density $\pi_i(\boldsymbol{\theta}_i)$ varies slowly and can be replaced by the constant $\pi_i(\hat{\boldsymbol{\theta}}_i)$. Then

$$\begin{aligned} q_i(\mathbf{x}) &\approx \pi_i(\hat{\boldsymbol{\theta}}_i) L_i \int \exp\{-(n/2)(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)' V_i^{-1}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)\} \mathrm{d}\boldsymbol{\theta}_i \\ &= \pi_i(\hat{\boldsymbol{\theta}}_i) L_i n^{-p_i/2} (2\pi)^{p_i/2} |V_i|^{1/2}, \end{aligned} \tag{4}$$

where p_i is the number of elements in $\boldsymbol{\theta}_i$.

This expansion is typically used to show that the posterior may be approximated by the $N(\hat{\boldsymbol{\theta}}_i, n^{-1} V_i)$ distribution. It obviously depends on standard regularity conditions, and those same conditions will be assumed for all asymptotic arguments in this paper. For a rigorous development of this approach to characterizing the asymptotic behaviour of Bayes factors, see Gelfand and Dey (1994). For model comparison, the Bayes factor is asymptotically given by the ratio of terms (4), and therefore

$$-2 \log B(\mathbf{x}) \approx -2 \log l + (p_1 - p_2) \log n + a, \tag{5}$$

where $l = L_1 / L_2$ is the classical likelihood ratio and

$$a = -\log \left\{ \frac{\pi_1(\hat{\boldsymbol{\theta}}_1)(2\pi)^{p_1/2} |V_1|^{1/2}}{\pi_2(\hat{\boldsymbol{\theta}}_2)(2\pi)^{p_2/2} |V_2|^{1/2}} \right\}$$

is $O(1)$. Ignoring a (or setting $a = 0$) produces the Schwarz (1978) test criterion $-2 \log l + (p_1 - p_2) \log n$. This, like Akaike's information criterion (Akaike, 1973)

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

of $-2\log l+2(p_1-p_2)$, adjusts the classical likelihood ratio criterion to favour more strongly the model with fewer parameters. The behaviour of these criteria can be examined by further asymptotic arguments.

Under regularity conditions the asymptotic sampling distribution of $-2\log l$ is found in classical statistics to be a non-central χ^2-distribution. See for example Stuart and Ord (1991), paragraph 23.7. Letting model 1 be the simpler model in a nested situation, write $\boldsymbol{\theta}_2=(\boldsymbol{\theta}_1,\ \phi)$ where $\phi=\phi_0$, a constant, under model 1. Then $-2\log l$ is asymptotically a non-central χ^2-variable with p_2-p_1 degrees of freedom and non-centrality parameter

$$\lambda=n(\phi-\phi_0)'\,V_\phi^{-1}(\phi-\phi_0),$$

where V_ϕ derives from the information matrix of a single observation. The expectation of $-2\log l$ is therefore asymptotically $p_2-p_1+\lambda$. If $\phi=\phi_0$, so that model 1 is true, $\lambda=0$ and the likelihood ratio criterion $-2\log l$ is $O(1)$, whereas if model 2 is true it is $O(n)$. The same is true of the Akaike criterion since $2(p_1-p_2)$ is also $O(1)$, but the Bayes factor and Schwarz criterion have different asymptotic behaviour. Under model 1, $-2\log B(\mathbf{x})$ is $O(-\log n)$, whereas under model 2 it is still $O(n)$.

Bayesian inference is therefore consistent when comparing nested models. If model 1 is true, $B(\mathbf{x})\to\infty$ and the posterior probability of model 1 tends to 1. If model 2 is true, $B(\mathbf{x})\to 0$ and the posterior probability of model 2 tends to 1. Although these results have been developed on the basis of independent and identically distributed (IID) observations, essentially the same asymptotics apply under increasing numbers of observations in linear models when the design matrix repeats cyclically or its rows are sampled at random. The development has, however, assumed proper priors $\pi_i(\boldsymbol{\theta}_i)$. It applies also to improper priors if the ratio c_1/c_2 is specified in advance, for instance by using the Spiegelhalter and Smith method. In general, any attempt to resolve the problem of improper priors, if it is to achieve the same consistency property, must have appropriate asymptotic behaviour.

The classical criterion $-2\log l$ and the Akaike variant would not produce consistency if used as Bayes factors. Under model 1, the posterior probability of model 1 would tend to a value less than 1. This reflects the classical hypothesis testing method, in which there is always a probability of wrongly rejecting the null hypothesis (model 1). The inconsistent behaviour of hypothesis tests as measures of evidence for or against a null hypothesis is emphasized by Berger and Delampady (1987) and Berger and Sellke (1987).

2. PARTIAL BAYES FACTORS

2.1. *Training Samples*

Another approach to improper priors makes use of a training sample. Divide the data into two parts, $\mathbf{x}=(\mathbf{y},\ \mathbf{z})$. The first part $\mathbf{y}$ will be used as a training sample to provide information about $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$, and the second part $\mathbf{z}$ will be used for model comparison. In the first step, $\mathbf{y}$ is used to obtain posterior distributions $\pi_i(\boldsymbol{\theta}_i|\mathbf{y})$ as in equation (2). Now taking these as prior distributions the remaining data $\mathbf{z}$ are used to compute a Bayes factor

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$B(\mathbf{z}|\mathbf{y}) = \frac{q_1(\mathbf{z}|\mathbf{y})}{q_2(\mathbf{z}|\mathbf{y})} = \frac{\int \pi_1(\boldsymbol{\theta}_1|\mathbf{y}) f_1(\mathbf{z}|\boldsymbol{\theta}_1, \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2|\mathbf{y}) f_2(\mathbf{z}|\boldsymbol{\theta}_2, \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_2}. \tag{6}$$

Any unspecified constants or impropriety in the priors $\pi_i(\theta_i)$ will have been removed in the first step, so that equation (6) is a well-defined Bayes factor based on proper priors $\pi_i(\theta_i|\mathbf{y})$.

$B(\mathbf{z}|\mathbf{y})$ will be referred to as a *partial* Bayes factor, as it is based on part of the data. It is also expressible as part of the full Bayes factor, since it is easy to demonstrate that

$$q_i(\mathbf{z}|\mathbf{y}) = \frac{\int \pi_i(\boldsymbol{\theta}_i) f_i(\mathbf{x}|\boldsymbol{\theta}_i)\, \mathrm{d}\boldsymbol{\theta}_i}{\int \pi_i(\boldsymbol{\theta}_i) f_i(\mathbf{y}|\boldsymbol{\theta}_i)\, \mathrm{d}\boldsymbol{\theta}_i} = \frac{q_i(\mathbf{x})}{q_i(\mathbf{y})}, \tag{7}$$

and therefore that

$$B(\mathbf{x}) = B(\mathbf{y})\, B(\mathbf{z}|\mathbf{y}). \tag{8}$$

This is just a consequence of the coherence of Bayes's theorem under sequential updating. Any indeterminacy in the full Bayes factor $B(\mathbf{x})$ applies also to $B(\mathbf{y})$ in equation (8), leaving the partial Bayes factor unaffected. Partial Bayes factors have been suggested by several previous researchers. The first reference seems to be Lempers (1971), chapter 6, who used half of the data as a training sample.

More recently, O'Hagan (1991) suggested using a proportion b of the data for training, whereas Berger and Pericchi (1993) advocated using a training sample of minimal size. Berger and Pericchi's suggestion is not affected by the discussion in Section 1.2 of difficulties with defining a minimal experiment. Their primary motivation is to devote as little of the data to training as necessary, to leave as much as possible for model comparison. If there were any debate about whether, for instance, two or three observations comprised a minimal size, then a training sample of 3 would be used. And, since the training sample is based on actual data, there is no difficulty over which values to use for regressor or design variables.

There is, however, a difficulty with all methods using partial Bayes factors about how to select the training sample from the data. With n observations, there are $\binom{n}{m}$ ways to choose a training sample of size m. Berger and Pericchi (1993) proposed using all possible training samples (which is feasible when the minimal sample size is sufficiently small) and averaging the resulting Bayes factors. They call such an average an *intrinsic* Bayes factor. However, it is not obvious how to average the factors. Berger and Pericchi considered both arithmetic and geometric mean forms of intrinsic Bayes factor. When the number of possible training samples of minimal size is large, they suggest averaging the partial Bayes factors from a random sample from this collection of possible training samples.

The primary purpose of this paper is to propose a simple alternative to averaging over many different selections of $\mathbf{y}$. This is the fractional Bayes factor (FBF) defined in Section 2.3.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

2.2. *Asymptotics of Partial Bayes Factors*

If $\mathbf{y}$ is fixed, and equation (4) is applied to the numerator of equation (7), exactly the same asymptotics will apply as in equation (5), with $q_i(\mathbf{y})$ now absorbed into the $O(1)$ term a. In particular, this is true when $\mathbf{y}$ is a minimal training sample, and the same will hold through the process of averaging Bayes factors from all possible minimal training samples. The intrinsic Bayes factor will yield consistent posterior inference, even starting from improper prior information.

Different asymptotics apply if m, the training sample size, also tends to ∞. Now applying equation (4) to both numerator and denominator gives

$$q_i(\mathbf{z}\,|\,\mathbf{y}) \approx \frac{\pi_i(\hat{\boldsymbol{\theta}}_i)\, L_i |\, V_i|^{1/2}\, n^{-p_i/2}}{\pi_i\{\hat{\boldsymbol{\theta}}_i(\mathbf{y})\} L_i(\mathbf{y}) |\, V_i(\mathbf{y})|^{1/2}\, m^{-p_i/2}}, \tag{9}$$

where adding $\mathbf{y}$ in parentheses denotes calculation of a quantity from the training sample. For instance, $\hat{\boldsymbol{\theta}}_i(\mathbf{y})$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i$ from data $\mathbf{y}$, whereas $\hat{\boldsymbol{\theta}}_i$ as before is based on the full data $\mathbf{x}$. Now equation (5) is replaced in general by

$$-2\log B(\mathbf{z}\,|\,\mathbf{y}) \approx -2\log l + 2\log l(\mathbf{y}) + (p_1 - p_2)(\log n - \log m) + O(1), \tag{10}$$

where l and $l(\mathbf{y})$ are the likelihood ratios from the full data and the training sample respectively. The asymptotic sampling distributions and expectations of $-2\log l$ and $-2\log l(\mathbf{y})$ can now be applied to obtain the behaviour of the partial Bayes factor generally.

If model 2 is true, $-2\log B(\mathbf{z}\,|\,\mathbf{y})$ is asymptotically $O(n-m)$, whereas if model 1 is true it becomes $O(\log m - \log n)$. Posterior probabilities will consistently choose the right model if n/m tends to ∞.

2.3. *Fractional Bayes Factors*

To avoid the arbitrariness of choosing a particular $\mathbf{y}$, or having to consider all possible subsets of a given size, a simplified form of partial Bayes factor may be defined as follows. Let $b = m/n$. If both m and n are large, the likelihood $f_i(\mathbf{y}\,|\,\boldsymbol{\theta}_i)$ based only on the training sample $\mathbf{y}$ will approximate to the full likelihood $f_i(\mathbf{x}\,|\,\boldsymbol{\theta}_i)$ raised to the power b. By analogy with equations (6) and (7) this motivates the alternative definition

$$B_b(\mathbf{x}) = q_1(b,\, \mathbf{x})/q_2(b,\, \mathbf{x}), \tag{11}$$

where

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

If $\pi_i(\boldsymbol{\theta}_i)$ has the improper form (1), the indeterminate constant c_i cancels out, leaving

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$q_i(b\,|\,\mathbf{x}) = \frac{\int h_i(\boldsymbol{\theta}_i)\, f_i(\mathbf{x}\,|\,\boldsymbol{\theta}_i)\, \mathrm{d}\boldsymbol{\theta}_i}{\int h_i(\boldsymbol{\theta}_i)\, f_i(\mathbf{x}\,|\,\boldsymbol{\theta}_i)^b\, \mathrm{d}\boldsymbol{\theta}_i}. \tag{13}$$

$B_b(\mathbf{x})$ will be referred to as an FBF. Although this definition is motivated asymptotically, it is proposed as an alternative form of partial Bayes factor even when m and n are not large. However, it is clearly asymptotically equivalent to the original definition (6), equations (9) and (10) becoming

$$q_i(b,\, \mathbf{x}) \approx L_i^{1-b}\, b^{p_i/2},$$

since the approximation entails $\hat{\boldsymbol{\theta}}_i(\mathbf{y}) = \hat{\boldsymbol{\theta}}_i$ and $V_i(\mathbf{y}) = V_i$, and

$$-2\log B_b(\mathbf{x}) \approx (1-b)\{-2\log l + (p_1 - p_2)\, c(b)\}, \tag{14}$$

with $c(b) = -\log b/(1-b)$. Equation (14) was given by O'Hagan (1991) as an asymptotic result for fixed b. However, consistency will require $b \to 0$ and hence $c(b) \to \infty$, as $n \to \infty$.

Equation (12) could be generalized further to $q_i(a,\, b,\, \mathbf{x}) = I_a(\mathbf{x})/I_b(\mathbf{x})$, where $I_b(\mathbf{x})$ is the denominator of equation (13) and $I_a(\mathbf{x})$ is the same with a replacing b. $B_{a,b}(\mathbf{x})$ is then defined analogously. The full Bayes factor $B(\mathbf{x})$ is $B_{1,0}(\mathbf{x})$. The FBF $B_b(\mathbf{x})$ is $B_{1,b}(\mathbf{x})$. We can now also include Aitkin's (1991) posterior Bayes factor $B_{2,1}(\mathbf{x})$. In place of equation (14) we have

$$-2\log B_{a,b}(\mathbf{x}) \approx (a-b)\{-2\log l + (p_1 - p_2)\, c(a,\, b)\}$$

with $c(a,\, b) = (\log a - \log b)/(a-b)$. Posterior probabilities derived from the posterior Bayes factor are not consistent because it gives $c(a,\, b) = \log 2$, which therefore does not tend to ∞ with the sample size.

3. SOME EXAMPLES

This section presents some simple examples to illustrate the derivation of FBFs, and to show their versatility.

3.1. *Testing Normal Mean*

Perhaps the simplest of all examples is as follows. Under model 1, $x_1, x_2, \ldots, x_n$ are IID $N(\theta_0,\, 1)$, whereas under model 2 they are IID $N(\theta,\, 1)$ and a uniform prior $h_2(\theta) = 1$ is assumed. Under model 1 there are no unknown parameters, and equation (13) reduces to

$$\begin{aligned} q_1(b,\, \mathbf{x}) &= f_1(\mathbf{x})/f_1(\mathbf{x})^b = f_1(\mathbf{x})^{1-b} \\ &= \left[(2\pi)^{-n/2} \exp\left\{-\sum_j (x_j - \theta_0)^2/2\right\}\right]^{1-b} \\ &= (2\pi)^{-n(1-b)/2} \exp\left\{-(1-b)\sum_j (x_j - \theta_0)^2/2\right\}. \end{aligned}$$

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

Under model 2, equation (13) has denominator

$$\int h_2(\theta) f_2(\mathbf{x}\,|\,\theta)^b \,\mathrm{d}\theta = (2\pi)^{-nb/2} \int \exp\left\{ -b \sum_j (x_j - \theta)^2/2 \right\} \mathrm{d}\theta$$

$$= (2\pi)^{-nb/2} \exp\left\{ -b \sum_j (x_j - \bar{x})^2/2 \right\} \int \exp\{-nb(\theta - \bar{x})^2/2\} \,\mathrm{d}\theta$$

$$= (2\pi)^{-nb/2} \exp\left\{ -b \sum_j (x_j - \bar{x})^2/2 \right\} (2\pi/nb)^{1/2}. \tag{15}$$

Hence

$$q_2(b,\ \mathbf{x}) = (2\pi)^{-n(1-b)/2} \exp\left\{ -(1-b) \sum_j (x_j - \bar{x})^2/2 \right\} b^{1/2} \tag{16}$$

and

$$B_b(\mathbf{x}) = \exp\left[-(1-b) \left\{ \sum_j (x_j - \theta_0)^2 - \sum_j (x_j - \bar{x})^2 \right\} /2 \right] b^{-1/2}$$

$$= \exp\{-n(1-b)(\bar{x} - \theta_0)^2/2\} b^{-1/2}. \tag{17}$$

In this case

$$-2 \log B_b(\mathbf{x}) = n(1-b)(\bar{x} - \theta_0)^2 + \log b \tag{18}$$

and equation (14) is exact. The FBF clearly has the expected property of decreasing as $\bar{x}$ moves away from θ_0. It is maximized at $b^{-1/2}$ when $\bar{x} = \theta_0$. Under model 1, $\bar{x}$ will tend in probability to θ_0. Then, provided that $b \to 0$ as stipulated in Section 2.3, $B_b(\mathbf{x}) \to \infty$ with probability 1. Under model 2, $B_b(\mathbf{x}) \to 0$ with probability 1.

In practice it is important to allow the variance to be unknown under both models. This is simply a special case of the general problem of comparing two linear models, which is considered in Section 4.

3.2. *Non-nested Example*

The FBF deals with non-nested models equally easily. As a simple example consider the problem in which IID observations $x_1, x_2, \ldots, x_n$ are distributed as $N(\theta_1, 1)$ under model 1, or as $N(0, \theta_2)$ under model 2. The standard non-informative priors are $h_1(\theta_1) = 1$, $h_2(\theta_2) = \theta_2^{-1}$. Now $q_1(b, \mathbf{x})$ has already been derived as equation (16) in the previous example, and we find

$$q_2(b,\ \mathbf{x}) = \left(\pi \sum_j x_j^2 \right)^{-n(1-b)/2} b^{bn/2} \frac{\Gamma(n/2)}{\Gamma(bn/2)}.$$

Therefore

$$B_b(\mathbf{x}) = k \left(\sum_j x_j^2 \right)^{n(1-b)/2} \exp\left\{ -(1-b) \sum_j (x_j - \bar{x})^2/2 \right\},$$

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

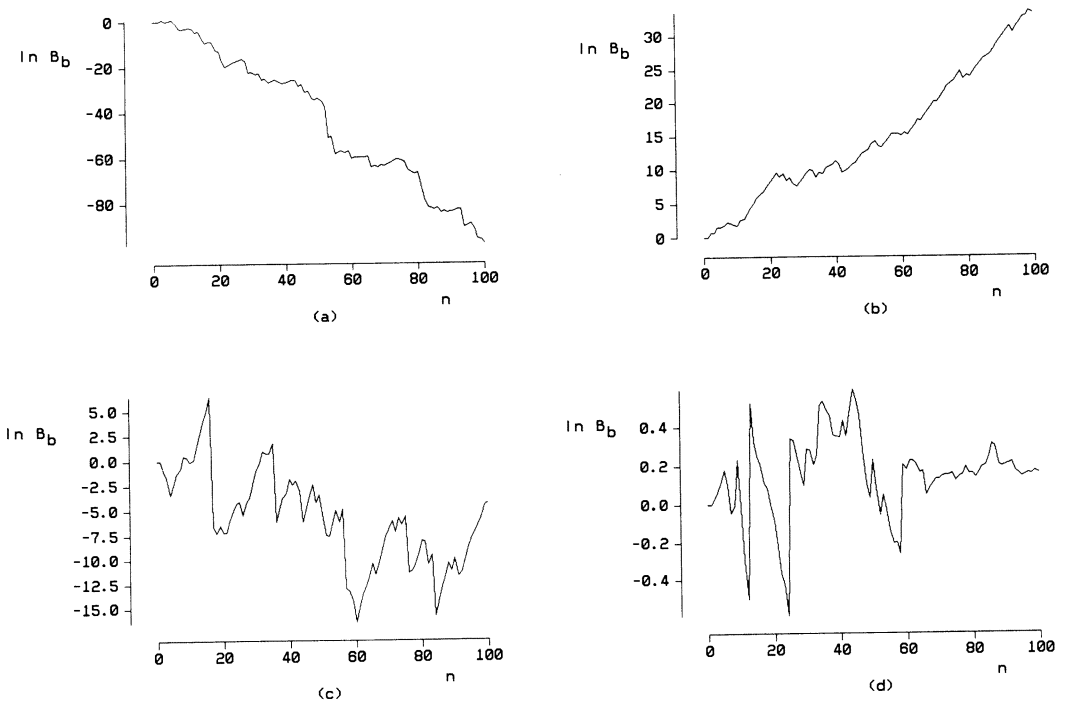


Fig. 1. Cumulative $\log B_b(\mathbf{x})$ for samples from (a) $N(0, 5)$, (b) $N(1, 1)$, (c) $N(2, 3)$ and (d) $N(0, 1)$ distributions

where

$$k = 2^{-n(1-b)/2} b^{(1-bn)/2} \frac{\Gamma(bn/2)}{\Gamma(n/2)}.$$

Fig. 1 shows the performance of $\log B_b(\mathbf{x})$ in four simulated data sets, each of 100 observations. In each case, b has been set to n^{-1}. Since $m = 1$ is the minimal training sample size to estimate the one parameter in each model, $b = n^{-1}$ is a kind of minimal value. The choice of b is considered more fully in Section 6. The first data set is generated from the $N(0, 5)$ distribution, so that model 2 is the correct model. Then we see $\log B_b(\mathbf{x})$ decline steadily towards $-\infty$ as more data are gathered. Data set (b) is generated from the $N(1, 1)$ distribution. Then the correct model is model 1 and $\log B_b(\mathbf{x})$ climbs steadily towards ∞. In both of these cases, 100 observations are more than enough to provide conclusive evidence in favour of the correct model.

The third case is of data from the $N(2, 3)$ distribution so that neither model is correct. Now $\log B_b(\mathbf{x})$ behaves quite erratically, with the data unable to choose reliably which of $N(\theta_1, 1)$ or $N(0, \theta_2)$ is 'more correct'. Finally, data set (d) is generated from the $N(0, 1)$ distribution, so that both models are correct. It is of course not possible to discriminate between the models, and $\log B_b(\mathbf{x})$ shows this by staying close to, and seemingly to converge to, 0. The behaviour is very different from the third data set, where the oscillations of $\log B_b(\mathbf{x})$ are very much larger and there is no apparent convergence.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

3.3. *Exponential versus Log-normal*

In this example, let $x_1, x_2, \ldots, x_n$ be IID exponential random variables with mean θ under model 1, and let their logarithms be IID $N(\mu, \sigma^2)$ under model 2. Again we adopt the standard non-informative priors $h_1(\theta) = \theta^{-1}$ and $h_2(\mu, \sigma^2) = \sigma^{-2}$. Under model 1 the likelihood is

$$f_1(\mathbf{x} \mid \theta) = \theta^{-n} \exp(-n\bar{x}/\theta)$$

and we find

$$q_1(b, x) = (n\bar{x})^{-n(1-b)} b^{bn} \Gamma(n)/\Gamma(bn).$$

Under model 2 the likelihood is

$$f_2(\mathbf{x} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \left(\prod_j x_j\right)^{-1} \exp\left\{-\sum_j (\log x_j - \mu)^2/2\sigma^2\right\}$$

and then

$$q_2(b, \mathbf{x}) = \left\{\pi \sum_j (\log x_j - \tilde{x})^2\right\}^{-n(1-b)/2} \left(\prod_j x_j\right)^{-(1-b)} b^{bn/2} \frac{\Gamma\{(n-1)/2\}}{\Gamma\{(bn-1)/2\}},$$

where $\tilde{x}$ is the mean of the $\log x_j$s.

The FBF $B_b(\mathbf{x}) = q_1(b, \mathbf{x})/q_2(b, \mathbf{x})$ is found to discriminate effectively in practice between the exponential and log-normal models, although this is a more difficult problem than the previous example. Fig. 2 plots $\log B_b(\mathbf{x})$ against sample size for

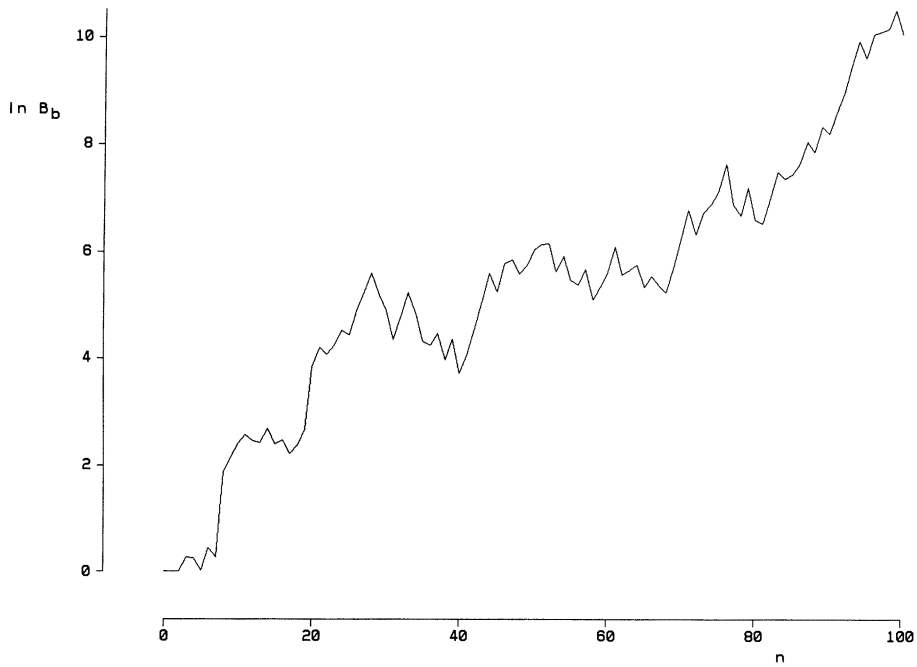


Fig. 2. Cumulative $\log B_b(\mathbf{x})$ for exponential(1) data

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

100 observations from the exponential distribution with mean $\theta = 1$. A minimal value for b of $2/n$ has been used again, based on $m = 2$ being a minimal training sample size for this problem. After 100 observations the Bayes factor is about $\exp(10) = 22000$, which quite conclusively points to the exponential model (although considerably less emphatically than the first two plots in Fig. 1).

4. LINEAR MODELS

The FBF is straightforward to derive for comparison between two normal linear models with conventional improper priors. For $i = 1, 2$, model i asserts that $\mathbf{x}$ is distributed as $N(\mathbf{Z}_i\boldsymbol{\beta}_i, \sigma_i^2\mathbf{I})$, where $\mathbf{Z}_i$ is an $n \times (p_i - 1)$ matrix of known coefficients, $\boldsymbol{\beta}_i$ is a $(p_i - 1)$-vector of unknown regression coefficients, $\mathbf{I}$ is the $n \times n$ identity matrix and σ_i^2 an unknown variance parameter. Then $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma_i^2)$ is p_i dimensional. The prior distribution under model i is the improper distribution specified by

$$h_i(\boldsymbol{\theta}_i) = (\sigma_i^2)^{-t_i},$$

for $i = 1, 2$. Setting $t_i = 1$, so that $h_i(\boldsymbol{\theta}_i) = \sigma_i^{-2}$, is generally used as an improper prior distribution for a linear model, but other values have also been proposed. In particular, the Jeffreys prior gives a t_i depending on p_i. Now

$$\int h_i(\boldsymbol{\theta}_i)\, f_i(\mathbf{x}|\boldsymbol{\theta}_i)^b \, \mathrm{d}\boldsymbol{\theta}_i = \pi^{-nb/2} \, |\mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i|^{-1/2} \times 2^{-r_i/2} \, b^{-(nb+p_i-r_i)/2} \, (S_i^2)^{-(nb-r_i)/2} \, \Gamma\{(nb-r_i)/2\}, \tag{19}$$

where $r_i = p_i - 2t_i + 1$ and S_i^2 is the residual sum of squares

$$S_i^2 = \mathbf{x}^{\mathrm{T}}\{\mathbf{I} - \mathbf{Z}_i(\mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i)^{-1}\mathbf{Z}_i^{\mathrm{T}}\}\mathbf{x}.$$

Substituting equation (19) into equation (12) gives

$$q_i(b, \mathbf{x}) = \pi^{-n(1-b)/2} \, b^{(nb+p_i-r_i)/2} \, (S_i^2)^{-n(1-b)/2} \, \frac{\Gamma\{(n-r_i)/2\}}{\Gamma\{(nb-r_i)/2\}}$$

and hence

$$B_b(\mathbf{x}) = \frac{\Gamma\{(n-r_1)/2\}\,\Gamma\{(nb-r_2)/2\}}{\Gamma\{(n-r_2)/2\}\,\Gamma\{(nb-r_1)/2\}} \, b^{t_1-t_2} \left(\frac{S_1^2}{S_2^2}\right)^{-n(1-b)/2}. \tag{20}$$

The FBF (20) is simple to compute. Unlike Bayes factors derived by Spiegelhalter and Smith (1982), it does not include the $|\mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i|^{-1/2}$-terms which appear in equation (19). This same feature is found in the criterion of de Vos (1993), who presented a way of obtaining the intrinsic Bayes factor for linear models, using a particular form of weighted averaging of the individual partial Bayes factors.

Notice that there is no need for the two linear models to be nested. In the nested case of testing a general linear hypothesis, the ratio of sums of squares S_1^2/S_2^2 is a function of the classical F-statistic. For example, consider the regression problem of Section 1.2 modified to give an unknown variance σ^2 under each model, with prior proportional to σ^{-2}. Then equation (20) becomes

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$B_b(\mathbf{x}) = \frac{\Gamma(n/2)\,\Gamma\{(nb-1)/2\}}{\Gamma\{(n-1)/2\}\,\Gamma(nb/2)}\{1+(n-1)^{-1}F\}^{-n(1-b)/2}$$

where F is the classical test statistic

$$F = (n-1)\hat{\beta}^2 \sum_j a_j^2 / \sum_j (x_j - a_j\hat{\beta})^2$$

and $\hat{\beta} = \Sigma_j a_j x_j / \Sigma_j a_j^2$.

5. SENSITIVITY

5.1. *Sensitivity to Prior—Example*

Consider again the example of Section 3.1, in which the data are IID $N(\theta, 1)$ under model 2, and model 1 asserts that $\theta = \theta_0$. Suppose that instead of the standard non-informative prior $h_2(\theta) = 1$ the specification $h_2(\theta) = \exp\theta$ is proposed. How sensitive are various Bayes factors to this change?

For the FBF, the denominator of equation (13) becomes

$$\int h_2(\theta) f_2(\mathbf{x}|\theta)^b \,\mathrm{d}\theta = (2\pi)^{-nb/2} \exp\left\{-b\sum_j (x_j-\bar{x})^2/2\right\} \int \exp(-Q/2)\,\mathrm{d}\theta,$$

where

$$Q = bn(\bar{x}-\theta)^2 - 2\theta = bn\{\theta - (\bar{x} + b^{-1}n^{-1})^2\} - 2\bar{x} - b^{-1}n^{-1}.$$

Therefore

$$\int h_2(\theta) f_2(\mathbf{x}|\theta)^b \,\mathrm{d}\theta = (2\pi)^{-nb/2} \exp\left\{-b\sum_j (x_j-\bar{x})^2/2\right\} (2\pi/bn)^{1/2} \exp\{\bar{x} + (2bn)^{-1}\}.$$

The last term here is new and results in the FBF being the previous expression (17) multiplied by $\exp\{(1-b)/2bn\}$. This multiplier therefore represents the sensitivity of $B_b(\mathbf{x})$ to the change from $h_2(\theta) = 1$ to $h_2(\theta) = \exp\theta$. Notice that it decreases as b increases. Increasing the training sample size reduces the sensitivity.

The same phenomenon is found with the partial Bayes factor $B(\mathbf{z}|\mathbf{y}) = q_1(\mathbf{z}|\mathbf{y})/q_2(\mathbf{z}|\mathbf{y})$, where $q_i(\mathbf{z}|\mathbf{y})$ is given by equation (7). For model 1, we find

$$q_1(\mathbf{z}|\mathbf{y}) = f_1(\mathbf{x})/f_1(\mathbf{y}) = (2\pi)^{(m-n)/2} \exp\left[\left\{\sum_j (y_j-\theta_0)^2 - \sum_j (x_j-\theta_0)^2\right\}/2\right],$$

and for model 2, by similar algebra to equation (15),

$$q_2(\mathbf{z}|\mathbf{y}) = (2\pi)^{(m-n)/2} \exp\left[\left\{\sum_j (y_j-\bar{y})^2 - \sum_j (x_j-\bar{x})^2\right\}/2\right](m/n)^{1/2}.$$

Therefore

$$B(\mathbf{z}|\mathbf{y}) = \exp[-\{n(\bar{x}-\theta_0)^2 - m(\bar{y}-\theta_0^2)\}/2](m/n)^{-1/2}. \tag{21}$$

With $h_2(\theta) = \exp\theta$, $B(\mathbf{z}|\mathbf{y})$ is found similarly to be equation (21) multiplied by $\exp\{\bar{y} - \bar{x} + (m^{-1} - n^{-1})/2\}$. This multiplier therefore represents the sensitivity of $B(\mathbf{z}|\mathbf{y})$ to the change of prior specification and naturally depends on the choice of a particular training sample $\mathbf{y}$. The maximum sensitivity is obtained by training samples

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

whose mean $\bar{y}$ differs as much as possible from $\bar{x}$, and this will also decrease as the size of training sample is increased.

The multiplier for $B(\mathbf{z}|\mathbf{y})$ is $\exp(\bar{y}-\bar{x})$ times that for $B_b(\mathbf{x})$, and this causes $B_b(\mathbf{x})$ to be less sensitive in general to the change in prior than $B(\mathbf{z}|\mathbf{y})$ is. Notice also that for a fixed training sample the sensitivity *increases* with sample size.

Now consider a *proper* $N(t, 1)$ prior distribution, $h(\theta) = \exp\{-(\theta-t)^2/2\}$. The FBF (11) is still well defined and is found to be

$$B_b(\mathbf{x}) = \exp\{-n(1-b)(\bar{x}-\theta_0)^2/2\}\{(nb+1)/(n+1)\}^{-1/2}\exp\{k(\bar{x}-t)^2/2\},$$

where $k = (1-b)n/(n+1)(nb+1)$ therefore represents the sensitivity of the Bayes factor to changes in the prior mean t. Notice again that this sensitivity decreases with b and increases with n. In this case the usual Bayes factor $B(\mathbf{x})$ is properly defined and corresponds to $b=0$. Therefore, even when proper prior distributions are specified and the usual Bayes factor is available for use, the FBF may be preferred because of its greater robustness to misspecification of the prior.

5.2. *General Sensitivity to Prior*

This example illustrates a much more general problem of sensitivity of Bayes factors to the prior distribution. It is well known that, as the amount of data increases, posterior inference about the parameters in a fixed model becomes increasingly less sensitive to the prior distribution. For a large sample, the likelihood $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ is negligible for $\boldsymbol{\theta}_i$ outside an $O(n^{-1/2})$ neighbourhood of $\hat{\boldsymbol{\theta}}_i$. Over that range, $\pi_i(\boldsymbol{\theta}_i)$ varies little and the posterior tends to the normalized likelihood, regardless of the prior. The key to this result is that the posterior depends only on the *relative* values of $\pi_i(\boldsymbol{\theta}_i)$ as $\boldsymbol{\theta}_i$ varies over the range supported by the likelihood. Absolute values of the prior density are irrelevant because of the normalization by $q_i(\mathbf{x})$. As the sample size increases, the range of $\boldsymbol{\theta}_i$-values supported by the likelihood shrinks, and for any smooth prior the relative variation over that range becomes negligible.

That is for inference within a single model. A very different situation obtains for model comparison. The terms $q_i(\mathbf{x})$ in the numerator and denominator of $B(\mathbf{x})$ depend on the average values of $\pi_i(\boldsymbol{\theta}_i)$ over the range supported by the likelihood. It is not relative values but *absolute* values that matter now. Dependence on the prior does not disappear as the sample size increases. For the same reason, the indeterminate c_is in improper priors do not disappear.

Moreover, there is a general sense in which sensitivity of model comparison to the prior *increases* with sample size, as sensitivity of inference within a given model decreases. If the likelihood is relatively diffuse, the variation of the prior density cannot be great in average value, since $\pi_i(\boldsymbol{\theta}_i)$ is constrained to integrate to 1. When $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ concentrates on a narrow range of $\boldsymbol{\theta}_i$-values, however, a relatively mild perturbation of the prior could produce a much larger change in the average value over a small range, leading to greater sensitivity of $B(\mathbf{x})$. It is almost as if there were a law of conservation of sensitivity, and increasing sample size only transfers sensitivity from inference about $\boldsymbol{\theta}_i$ within model i to inference between models.

Use of partial Bayes factors reduces this sensitivity, as the example of Section 5.1 illustrates. Sensitivity of $B(\mathbf{z}|\mathbf{y})$ to the prior is now expressed as sensitivity to $\pi_i(\boldsymbol{\theta}_i|\mathbf{y})$. This is the posterior distribution of $\boldsymbol{\theta}_i$ after the training sample. As

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

discussed at the beginning of this section, the more observations there are in the training sample, the less sensitive will $\pi_i(\theta_i|\mathbf{y})$ be to variation of the prior $\pi_i(\theta_i)$. Sensitivity of the model comparison is thereby also decreased. The intrinsic Bayes factor is based on a minimal training sample. For proper priors, the minimal training sample is no sample at all and there is no reduction in sensitivity. For improper priors, it derives the minimal benefit of reduced sensitivity. The preceding example also shows that the FBF is in general less sensitive than partial Bayes factors.

The goal instead should be to make the best possible use of this property of partial Bayes factors. If a large amount of data is available, we can afford to devote a relatively large amount of data to the training sample to achieve robustness. As $n \to \infty$, complete robustness of the partial Bayes factor to the prior is obtained by letting m also tend to ∞. As seen in Section 2.2, if $m \to \infty$ as fast as n, so that $m/n = b$ is constant, the partial Bayes factor will not produce consistent comparisons between nested models. But if, for instance, m increases as $O(\log n)$, then both asymptotic robustness and consistency can be achieved. The same benefits are also achieved by FBFs with $b = O(n^{-1} \log n)$.

Although the present paper is not primarily concerned with model comparison when *proper* priors are available, it is my belief that FBFs could also be beneficial there in reducing sensitivity to the prior.

5.3. *Sensitivity to Outliers*

Use of the FBF $B_b(\mathbf{x})$ provides another way of reducing sensitivity. If the partial Bayes factor $B(\mathbf{z}|\mathbf{y})$ is computed over all (or many of) the possible ways of dividing the sample into $\mathbf{z}$ and $\mathbf{y}$, then its value can vary greatly, particularly if there are outliers in the data. In general, some observations may be highly influential for the parameters of one model but not the other. In nested models, this arises with observations that are highly influential for parameters which are missing in (or set to specific hypothesized values by) the simpler model. Then there will be some divisions of the data into $\mathbf{y}$ and $\mathbf{z}$ which produce 'outlying' partial Bayes factors $B(\mathbf{z}|\mathbf{y})$. The sensitivity is greatest when $\mathbf{y}$ is minimal, as in the intrinsic Bayes factor. It disappears if $B_b(\mathbf{x})$ is used instead.

Consider again the simple example of Section 3.1. The partial Bayes factor $B(\mathbf{z}|\mathbf{y})$ was found in Section 5.1, equation (21). When the sample contains outliers, $\bar{y}$ can be far from $\bar{x}$ for some training samples, particularly if m is small. When computing an intrinsic Bayes factor, m is made as small as possible, and the outlying training samples can strongly influence the averaging process. In this example, the minimal training sample size is $m = 1$, and if the jth observation x_j is selected for training then the partial Bayes factor is $B(\mathbf{x}_{(j)}|x_j)$, where

$$\begin{aligned} -2\log B(\mathbf{x}_{(j)}|x_j) &= (n-1)(\bar{x}_{(j)}-\theta_0)^2 - \log(n-1) - \log\{1+(n-1)^{-1}\} \\ &\quad -\{1+(n-1)^{-1}\}^{-1}(\bar{x}_{(j)}-x_j)^2 \\ &= n(\bar{x}-\theta_0)^2 - (x_j-\theta_0)^2 - \log n. \end{aligned} \tag{22}$$

Even if the data support model 2 strongly, as measured by $\bar{x}$ being relatively far from θ_0, a single observation x_j far from θ_0 can produce a partial Bayes factor which actually favours model 1. Averaging these partial Bayes factors $B(\mathbf{x}_{(j)}|x_j)$ will yield an arithmetic mean form of intrinsic Bayes factor $\bar{B}$, where

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$-2\log \bar{B} = n(\bar{x}-\theta_0)^2 + \bar{D} - \log n, \tag{23}$$

where

$$\bar{D} = -2\log n^{-1} \sum_{j=1}^{n} \exp\{(x_j-\theta_0)^2/2\}.$$

A single outlier can dominate in this sum and produce a highly misleading Bayes factor, as in the following example.

A much analysed data set with outliers is Darwin's data; see for instance Box and Tiao (1962). The data comprise the 15 observations 49, −67, 8, 16, 6, 23, 28, 41, 14, 56, 24, 75, 60, −48, 29. Dividing by 20 produces data that might reasonably be supposed to be from a normal distribution with mean 1 and variance 1, but with several outliers. The values of equation (22) range from −21.60 to −2.70, corresponding to Bayes factors ranging from 48968, very strongly favouring $\theta_0 = 1$, to 3.85 which represents very much weaker evidence in favour of that value. However, the very large Bayes factors all result from outliers in the original data. Expression (23) is −16.24, resulting in an intrinsic Bayes factor $\bar{B} = 3364$ which has been heavily influenced by the most extreme of the outliers. Berger and Pericchi (1993) are aware of this difficulty and propose that the arithmetic mean intrinsic Bayes factor should always be computed with the more complex of the two models in the numerator. In this case their factor is defined by first inverting the factors $B(\mathbf{x}_{(j)}|x_j)$ before averaging, and then inverting the average. This yields an intrinsic Bayes factor of 6.62, which is far less influenced by the outliers. Berger and Pericchi discuss the problem of determining a 'more complex' model in the case of non-nested models, and that of comparing three or more models, and suggest various modified factors.

Berger and Pericchi's alternative suggestion of geometric averaging implies averaging not $B(\mathbf{x}_{(j)}|x_j)$ but $-2\log B(\mathbf{x}_{(j)}|x_j)$, thereby obtaining

$$n(\bar{x}-\theta_0)^2 - n^{-1}\sum_{j=1}^{n}(x_j-\theta_0)^2 - \log n = (n-1)(\bar{x}-\theta_0)^2 - n^{-1}\sum_{j=1}^{n}(x_j-\bar{x})^2 - \log n, \tag{24}$$

but here the sample variance also affects unnaturally the model comparison. On the Darwin data equation (24) yields a value of −6.00 and a corresponding Bayes factor of 20.8.

For the FBF, set $b = n^{-1}$ as in previous examples to have the effect of a training sample of size 1. Then equation (18) becomes

$$-2\log B_b(\mathbf{x}) = (n-1)(\bar{x}-\theta_0)^2 - \log n. \tag{25}$$

The sensitivity to outliers, or to the sample variance, is eliminated. For the Darwin data, equation (25) is −2.68, corresponding to an FBF of 3.814 which is more conservative than any of the various intrinsic Bayes factors. In fact it corresponds quite closely to taking the most conservative value of all the possible partial Bayes factors $B(\mathbf{x}_{(j)}|x_j)$.

6. CHOICE OF b

The key question remaining in the use of FBFs is the choice of b. It may seem that the only achievement of this paper is to replace an arbitrary ratio c_1/c_2 in the

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

usual Bayes factor (3), or an arbitrary choice of imaginary experiment in the Spiegelhalter and Smith approach, by an arbitrary choice of b. Even if this were the case, there is progress because the arbitrariness of the ratio c_1/c_2, or of the imaginary experiment in some kinds of application, allows the Bayes factor to take any value at all from 0 to ∞. As b is varied from m_0/n, where m_0 is the minimal training sample size, to 1 the value of $B_b(x)$ is strictly bounded. It will, for instance, almost always lie on one side of 1 for all those values of b, and hence it will be clear whether the data favour model 1 or model 2.

Furthermore, Section 2.3 makes it clear that b should tend to 0 as $n \to \infty$, to achieve consistent model choice. This criterion is satisfied by the minimal value $b = m_0/n$, and this has been used in all the numerical examples of earlier sections. If robustness to misspecification of the prior or the models (as in the possibility of outliers) is not a serious concern, then $b = m_0/n$ is a natural choice. It makes maximal possible use of the data for model comparison.

If, however, robustness is a serious concern, then Section 5.2 suggests strongly that a larger value of b is advisable, and that perfect asymptotic robustness may be achieved by letting $nb \to \infty$ as $n \to \infty$. Consistent with $b \to 0$ and $nb \to \infty$, the cases $b = n^{-1} \log n$ and $b = n^{-1/2}$ are worthy of consideration. The first corresponds to a training sample size $m = \log n$, which increases very slowly with n. It makes some allowance for achieving robustness but in general keeps the training sample size very small. The second corresponds to a training sample size $m = \sqrt{n}$, which places much more stress on robustness.

Formally, then, I propose three ways to set b:

(a) $b = m_0/n$, when robustness is no concern,
(b) $b = n^{-1} \max\{m_0, \sqrt{n}\}$, when robustness is a serious concern, and
(c) $b = n^{-1} \max\{m_0, \log n\}$, as an intermediate option.

Figs 3 and 4 illustrate the effect of using the three different choices of b in practice. Fig. 3 corresponds to the first data set used in Fig. 1. The difference between the three choices of b is not large, and even with the square root choice the preference for model 2 is clear with very small samples. Fig. 4 uses the same data as Fig. 2. In this case the data are less strong in favouring the exponential model, and the differences between the three choices of b can be sufficiently substantial to have a noticeable effect on posterior probabilities of the two models.

7. COHERENCE

7.1. *Sufficiency*

Several questions can be raised concerning coherence of partial Bayes factors. First consider sufficiency. Coherent Bayesian inference is always a function of the sufficient statistics, a consequence of the likelihood principle. The FBF $B_b(\mathbf{x})$ is coherent in this sense, since it clearly depends only on the likelihood function. This is not true of $B(\mathbf{z} \mid \mathbf{y})$, as is shown by equation (22). In that example $\bar{x}$ is sufficient (under both models), but $B(\mathbf{x}_{(j)} \mid x_j)$ depends on x_j as well as $\bar{x}$. Even averaging to yield equation (24) produces an intrinsic Bayes factor depending on the sample variance as well as $\bar{x}$. Equation (25) confirms that $B_b(\mathbf{x})$ depends only on the sufficient statistics.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

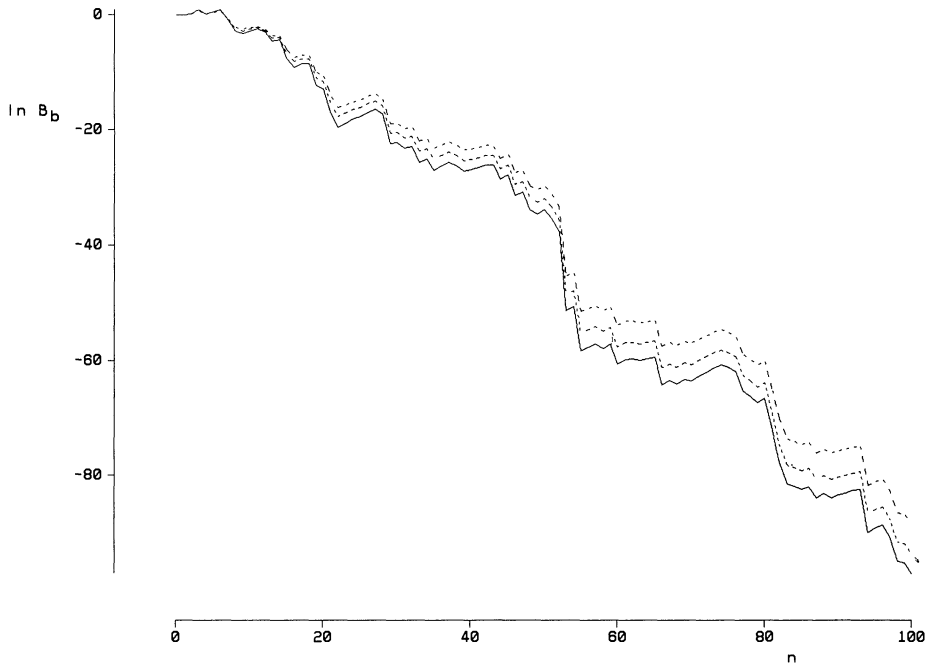


Fig. 3. Cumulative log $B_b(\mathbf{x})$ on $N(0, 5)$ data: ——, minimal b, · · · · ·, logarithmic b; ------, square root b

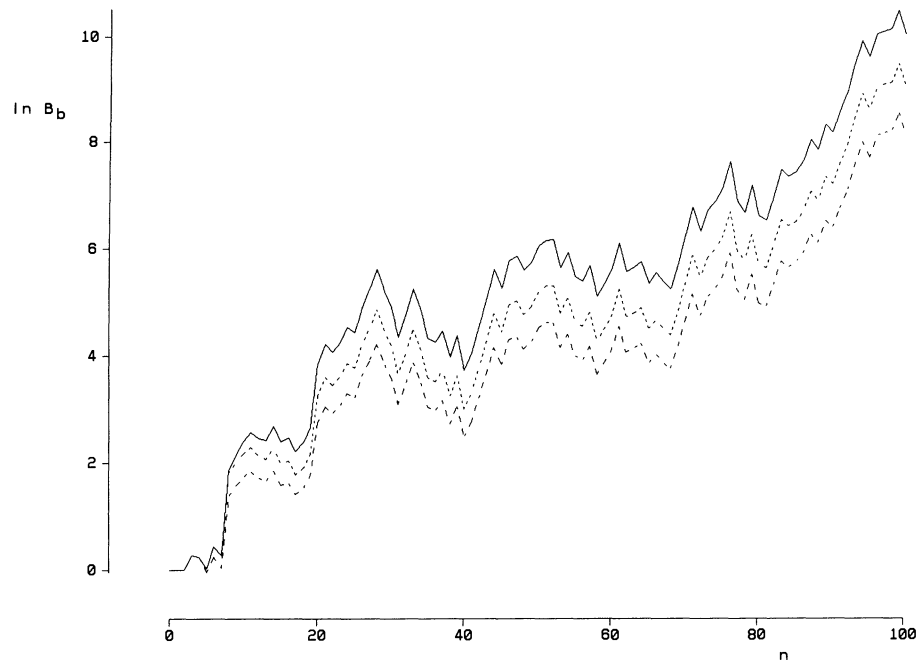


Fig. 4. Cumulative log $B_b(\mathbf{x})$ on exponential(1) data: ——, minimal b; · · · · ·, logarithmic b; ------, square root b

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

7.2. *Coherence for Given Sample*

After the training sample $\mathbf{y}$ has been observed, $\pi_i(\theta_i|\mathbf{y})$ is the correct posterior distribution, which then becomes the prior distribution for calculating a genuine Bayes factor $B(\mathbf{z}|\mathbf{y})$ based on the data $\mathbf{z}$. It differs from the full Bayes factor only in that it ignores $\mathbf{y}$ for model comparisons, discarding the term $B(\mathbf{y})$ in equation (8). It therefore does not use all the data for the model comparison stage (and this is the source of its dependence separately on the sufficient statistics from both $\mathbf{z}$ and $\mathbf{y}$, rather than the sufficient statistics of the full data $\mathbf{x}$). As long as $\mathbf{y}$ is not chosen deliberately to influence the value of the partial Bayes factor, it does not seem incoherent to ignore some of the data, merely potentially wasteful. This contrasts with the obvious non-coherence of Aitkin's (1991) posterior Bayes factor, where the full data $\mathbf{x}$ are used as a training sample and then reused for model comparison.

Averaging partial Bayes factors from different training samples is not coherent in this way. Whereas $B(\mathbf{z}|\mathbf{y})$ is a genuine Bayes factor, the averaging (as in the intrinsic Bayes factor) does not result in a single Bayes factor. Similarly, $B_b(\mathbf{x})$ is not a genuine Bayes factor, although the idea of using an idealized fraction of the data as a training sample seems sensible. If $bn \to \infty$, $B_b(\mathbf{x})$ is asymptotically equivalent to a single partial Bayes factor.

7.3. *Sequential Coherence*

Coherence is by no means assured if further data $\mathbf{x}^*$ become available. Again the partial Bayes factor with fixed $\mathbf{y}$ passes the test. As in equation (8),

$$B(\mathbf{z}, \mathbf{x}^*|\mathbf{y}) = B(\mathbf{z}|\mathbf{y})\, B(\mathbf{x}^*|\mathbf{z}, \mathbf{y}).$$

The new partial Bayes factor is the product of the partial Bayes factor $B(\mathbf{z}|\mathbf{y})$ and $B(\mathbf{x}^*|\mathbf{x})$. This is the same process as that by which a full Bayes factor would be updated sequentially from $B(\mathbf{x})$ to $B(\mathbf{x}, \mathbf{x}^*)$.

However, fixing $\mathbf{y}$ is arbitrary and can be highly sensitive to the original choice of $\mathbf{y}$ from data $\mathbf{x}$. In practice, Berger and Pericchi's (1993) approach of averaging many partial Bayes factors is obviously desirable to reduce that sensitivity (although Section 5.3 shows that this is an inadequate remedy). If the average is then updated by $B(\mathbf{x}^*|\mathbf{x})$, the result is an averaging of partial Bayes factors $B(\mathbf{z}, \mathbf{x}^*|\mathbf{y})$. There is coherence in this, except that all the training samples in the averaging are drawn from the original data $\mathbf{x}$. If *all* possible training samples are to be used, the average cannot be updated in this way, and many more terms must be included in the averaging.

The FBF cannot be updated coherently by multiplying by $B(\mathbf{x}^*|\mathbf{x})$ either. $B_b(\mathbf{x}, \mathbf{x}^*)$ is not derivable from $B_b(\mathbf{x})$, $\pi_i(\theta_i|\mathbf{x})$ and $f_i(\mathbf{x}^*|\theta_i, \mathbf{x})$ alone, and in practice would need to be calculated from scratch.

8. CONCLUSIONS

This paper has considered various properties of partial Bayes factors, and in particular of FBFs. In general, FBFs are preferred because of their greater robustness and their conformance with the likelihood principle.

Asymptotically, both consistency (selecting the correct model with probability 1 in nested problems) and complete robustness can be achieved by using $B_b(\mathbf{x})$ with $b \to 0$ but $bn \to \infty$ as $n \to \infty$.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

In finite samples, there is a trade-off between robustness, which increases with b, and discriminatory power, which decreases with b. An extreme position is represented by the FBF with minimal $b = m_0/n$, which may be appropriate when robustness to misspecification of the prior is not important. But where there is uncertainty over the specification of the prior a choice of $b = n^{-1} \log n$ or $b = n^{-1/2}$ is recommended.

Any practical application of partial Bayes factors or FBFs will be difficult to apply coherently in a sequential mode. This seems to be the only real drawback to their use.

ACKNOWLEDGEMENT

I would like to thank three referees for their helpful and insightful comments on an earlier version of this paper.

REFERENCES

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. R. Statist. Soc.* B, **53**, 111–142.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory*, pp. 267–281. Budapest: Akademia Kiadoó.

Berger, J. O. and Delampady, M. (1987) Testing precise hypotheses (with discussion). *Statist. Sci.*, **2**, 317–352.

Berger, J. O. and Pericchi, L. R. (1993) The intrinsic Bayes factor for model selection. *Technical Report 93-43C.* Department of Statistics, Purdue University, West Lafayette.

Berger, J. O. and Sellke, T. (1987) Testing of a point null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *J. Am. Statist. Ass.*, **82**, 112–139.

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

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B, **56**, 501–514.

Lempers, F. B. (1971) *Posterior Probabilities of Alternative Linear Models.* Rotterdam: University Press.

O'Hagan, A. (1991) Discussion on Posterior Bayes factors (by M. Aitkin). *J. R. Statist. Soc.* B, **53**, 136.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

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

Stuart, A. and Ord, J. K. (1991) *Kendall's Advanced Theory of Statistics*, 5th edn, vol. 2. London: Arnold.

de Vos, A. F. (1993) A fair comparison between regression models of different dimension. Submitted to *J. Econometr.*

DISCUSSION OF THE PAPER BY O'HAGAN

W. R. Gilks (Medical Research Council Biostatistics Unit, Cambridge): Professor O'Hagan has presented a delightfully clear account of the problem of Bayesian model choice with improper priors and has provided an elegant solution. I would like to make some practical points.

Scope of problem

The formal Bayesian machinery of model choice depends on the marginal likelihood $q_i(\mathbf{x})$ defined in Section 1.1. As the author notes, this quantity is not well defined for improper priors π_i. The use of proper but vague priors does little to alleviate the problem as $q_i(\mathbf{x})$ can then be very sensitive to the choice of π_i, as is evident from the first term in the Laplace approximation of equation (4). Thus we are naturally led to question the very enterprise of model choice. If the aim of statistical analysis is to predict (and some would argue that this is always the case), then there is no need to choose between models. Prediction can be better done on the basis of a panoply of models, as is demonstrated in Draper (1995). However, even here $q_i(\mathbf{x})$ must be evaluated, since the predictive distribution of future observations y is

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$$P(y \mid \mathbf{x}) = \frac{\sum_i P(y \mid \mathbf{x}, M_i)\, P(M_i)\, q_i(\mathbf{x})}{\sum_i P(M_i)\, q_i(\mathbf{x})},$$

in the notation of the paper. Thus, whenever more than one model is entertained, a robust form of $q_i(\mathbf{x})$ is required.

Solution

The robust form of $q(\mathbf{x})$ proposed in the paper is

$$q(b, \mathbf{x} \mid \tau) = \int f^{1-b}(\mathbf{x} \mid \boldsymbol{\theta})\, \pi_b(\boldsymbol{\theta} \mid \tau)\, \mathrm{d}\boldsymbol{\theta} = E_{\boldsymbol{\theta}}^{(b)}(f^{1-b}), \tag{26}$$

where $E_{\boldsymbol{\theta}}^{(b)}$ denotes expectation over the following distribution for $\boldsymbol{\theta}$:

$$\pi_b(\boldsymbol{\theta} \mid \tau) \propto f^b(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \tau). \tag{27}$$

Here I have introduced a parameter τ which I shall need below, and I have suppressed the subscript i, but otherwise this expression corresponds to that given in equation (12) with some algebraic rearrangement. I shall call $q(b, \mathbf{x} \mid \tau)$ a *fractional* marginal likelihood, and $\pi_b(\boldsymbol{\theta} \mid \tau)$ a *fractional posterior* since it has the form of a posterior distribution.

Optimal choice of b

In equation (26) some information about $\boldsymbol{\theta}$ has been transferred from the likelihood to the prior, so that the prior is no longer vague or in conflict with the data. This is similar in spirit to the partial and intrinsic Bayes factors. Professor O'Hagan usefully departs from tradition here by emphasizing that the amount of information transferred from likelihood to prior can be greater than required to make the prior proper. Transferring too little (setting b too small) will render $q(b, \mathbf{x} \mid \tau)$ sensitive to π; setting b too large will reduce the information available for model discrimination.

This suggests a lurking optimization problem. Professor O'Hagan has avoided making this explicit, preferring to give some informal advice on the choice of b, such as choosing $b = n^{-1} \log n$. To formalize the problem, I have above introduced the parameter τ to index a family of prior distributions, so that uncertainty about π is represented by uncertainty about τ. With a prior $\phi(\tau)$ for τ we may define a fractional posterior for τ:

$$\phi_b(\tau) \propto \phi(\tau) \int f^b(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \tau)\, \mathrm{d}\boldsymbol{\theta}.$$

Then we have

$$q(b, \mathbf{x}) = E_{\tau}^{(b)} E_{\boldsymbol{\theta}}^{(b)}(f^{1-b}) = \frac{\int f(\mathbf{x} \mid \boldsymbol{\theta}) \int \pi(\boldsymbol{\theta} \mid \tau)\, \phi(\tau)\, \mathrm{d}\tau\, \mathrm{d}\boldsymbol{\theta}}{\int f^b(\mathbf{x} \mid \boldsymbol{\theta}) \int \pi(\boldsymbol{\theta} \mid \tau)\, \phi(\tau)\, \mathrm{d}\tau\, \mathrm{d}\boldsymbol{\theta}} \tag{28}$$

where $E_{\tau}^{(b)}$ denotes expectation over $\phi_b(\tau)$. Equation (28) is the fractional marginal likelihood that would have resulted if full uncertainty about the prior had been acknowledged. However, the point of introducing τ was not because we necessarily believe in it: it is merely a device for measuring the sensitivity of $q(b, \mathbf{x} \mid \tau)$ to the prior, which might be defined as

$$\mathrm{var}_{\tau}^{(b)}\{q(b, \mathbf{x} \mid \tau)\} = \mathrm{var}_{\tau}^{(b)}\{E_{\boldsymbol{\theta}}^{(b)}(f^{1-b})\}.$$

The information available for model discrimination might then be represented as

$$E_{\tau}^{(b)}\{\mathrm{var}_{\boldsymbol{\theta}}^{(b)}(f^{1-b})\}.$$

This suggests choosing b to maximize the log-variance ratio:

$$\log[E_{\tau}^{(b)}\{\mathrm{var}_{\boldsymbol{\theta}}^{(b)}(f^{1-b})\}] - \log[\mathrm{var}_{\tau}^{(b)}\{E_{\boldsymbol{\theta}}^{(b)}(f^{1-b})\}], \tag{29}$$

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

or the sum of contributions of the form of expression (29) over all models of interest. Although this might provide a theoretical framework for deciding on b, it leaves open the question about how to choose the priors π and ϕ. Furthermore, the calculations required by expression (29) are difficult in even simple situations. One possible way forwards might be to use an ϵ-contamination model for $\pi(\boldsymbol{\theta} \mid \tau)$ (see for example Wasserman (1992)).

Hierarchical models

In the paper, fractional Bayes factors are motivated and discussed solely for the exchangeable data model

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} g(x_j \mid \boldsymbol{\theta}).$$

Though this class of models encompasses a great many applications, it is uncomfortably restrictive in view of the breadth of models which can be routinely handled by using Markov chain Monte Carlo (MCMC) methods. For example, how should the fractional Bayes factor be defined for the following hierarchical model:

$$f(\mathbf{x}, y \mid \boldsymbol{\theta}) = \prod_{j=1}^{m} g(x_j \mid \boldsymbol{\theta}) \prod_{k=1}^{n} h(y_{jk} \mid \boldsymbol{\theta}, x_j)?$$

Here it would seem unreasonable to apply the same value of b to both levels in the hierarchy. A natural extension of the methodology would be to define the fractional marginal likelihood as

$$q(a, b, \mathbf{x}) = \frac{\int g(\mathbf{x} \mid \boldsymbol{\theta})\, h(y \mid \mathbf{x}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}{\int g^a(\mathbf{x} \mid \boldsymbol{\theta})\, h^b(y \mid \mathbf{x}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}.$$

For yet more complex models, it could become increasingly unclear how to 'fractionate' the likelihood. Some guidance would be welcome.

Calculation of fractional marginal likelihood

The analytical approach used in the simple examples in the paper is clearly not an option for even moderately more complex applications, such as generalized linear models. Fortunately, a way out is provided by the MCMC method. Noting from equation (26) that $q(b, \mathbf{x} \mid \tau)$ is an expectation of f^{1-b}, we could perform MCMC iterations on the fractional posterior (27), estimating $q(b, \mathbf{x} \mid \tau)$ as the average of the values of f^{1-b} in the MCMC-generated samples. This could be expected to work well, since the fractional posterior will tend to deliver samples in regions of high likelihood. Moreover, for most likelihoods encountered in practice, simulation from the fractional posterior should be no more difficult than simulation from the full posterior (the usual role of the MCMC method). Note that the analogy of this approach for the standard non-fractional Bayes factor would generally not work since it would involve sampling from the prior, which tends to deliver few samples in regions of high likelihood (Newton and Raftery, 1994).

Other issues of model determination

Of course there are many other issues in model determination apart from the calculation of marginal likelihoods. With the MCMC method, residuals of various forms can be calculated and plotted and departures from the model tested, perhaps by using Bayesian p-values (Gelman *et al.*, 1994), and cross-validatory ideas can also be employed (Gelfand *et al.*, 1992).

Finally, I would like to congratulate Professor O'Hagan on a thoroughly interesting and useful paper. I propose the vote of thanks.

A. F. M. Smith (Imperial College of Science, Technology and Medicine, London): The problem that this paper addresses is that of choosing between models. But what is a model?

From the de Finetti perspective, a model is essentially a predictive machine for observable quantities, with the predictive density form

$$q(\mathbf{x}) = \int \pi(\boldsymbol{\theta}) f(\mathbf{x} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

justified by some kind of representation theorem. As we acquire data $x_1, x_2, \ldots$, the initial pair $\{f(\ |\theta), \pi(\theta)\}$ is replaced, successively, by $\{f(\ |\theta), \pi(\theta|x_1)\}$, $\{f(\ |\theta), \pi(\theta|x_1, x_2)\}$ and so on. The predictive model for the future is defined by the current pair and only acquires a predictive capability once $\pi(\theta|\ )$ is *proper*. Bayesians usually emphasize the dangers of focusing on a likelihood in the absence of a prior. But Professor O'Hagan seems to be regarding $f(\ |\theta)$ as 'the model', with $\pi(\theta)$ an irritating extra.

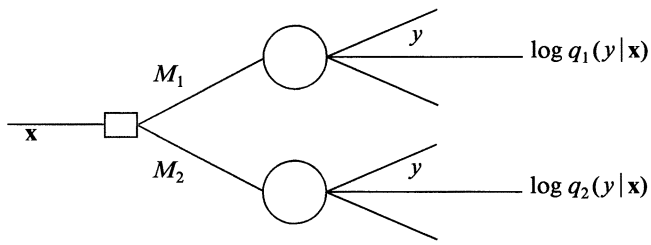
Moreover, with M_i, $i=1, 2$, denoting alternative models, the Bayes factor is defined to be

$$B_{12}(\mathbf{x}) = \frac{P(M_1|\mathbf{x})}{P(M_2|\mathbf{x})} \Big/ \frac{P(M_1)}{P(M_2)},$$

which implicitly assumes that it makes sense to attach probabilistic beliefs to models. If we take the view that models are just simplified artefacts for helping to structure the way that we think, the status of $P(M_i)$, $P(M_i|\mathbf{x})$, $i=1, 2$, is, to say the least, debatable.

So, is there any role for the Bayes factor in model choice problems? And, if so, which of the many varieties of Bayes factor on offer should we use?

Let us consider a simple decision problem. Given $\mathbf{x}=(x_1, \ldots, x_n)$ we want to predict $y=x_{n+1}$. For this, two 'off-the-shelf' models, M_1 and M_2, are available (computer packaged!) providing predictive densities $q_i(y|\mathbf{x})=q(y|\mathbf{x}, M_i)$, $i=1, 2$. But, of course, neither M_1 nor M_2 represents our *actual* beliefs. These we denote by $q_A(y|\mathbf{x})$, although (through ignorance or indolence) they remain unformulated. If we use the logarithmic utility function to score our predictions, we arrive at the following decision tree representation of the problem:



The Bayesian solution is to choose M_1 if

$$\int \log\left\{\frac{q_1(y|\mathbf{x})}{q_2(y|\mathbf{x})}\right\} q_A(y|\mathbf{x})\, dy > 0.$$

But, how do we evaluate the integral (since q_A is unspecified)? First, note that $(\mathbf{x}, y)=[(x_1, \ldots, x_n), x_{n+1}]$ can be mimicked by $(\mathbf{x}_{n-1}(j), x_j)$ for any $j=1, \ldots, n$, where $\mathbf{x}_{n-1}(j)=\mathbf{x}\setminus\{x_j\}$. Then, in the case of exchangeable x_is, a Monte Carlo approximation of the integral is available in the form

$$\frac{1}{K}\sum_{j=1}^{K} \log\left[\frac{q_1\{x_j|\mathbf{x}_{n-1}(j)\}}{q_2\{x_j|\mathbf{x}_{n-1}(j)\}}\right] = \frac{1}{K}\sum_{j=1}^{K} \log B_{12}\{x_j, \mathbf{x}_{n-1}(j)\},$$

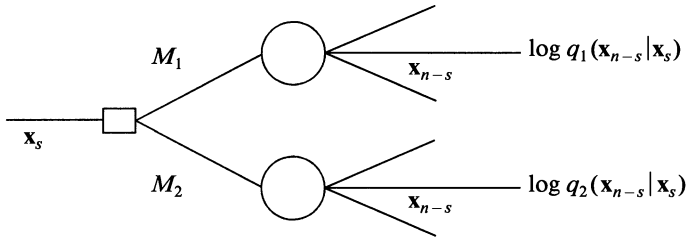
say, based on K random drawings from the possible partitions of $\mathbf{x}$. The (approximated) solution to the model choice problem is thus to choose M_1 if

$$\prod_{j=1}^{K} [B_{12}\{x_j, \mathbf{x}_{n-1}(j)\}]^{1/K} > 1.$$

We have, therefore, found a role for the Bayes factor, without implicit reference to prior or posterior probabilities on models; or, more precisely, a role for the geometric mean of a version of what Berger and Pericchi (1993) called 'intrinsic' Bayes factors.

The above illustration ('predicting the future') was based on taking the pair $\{f(x_{n+1}|\theta), \pi(\theta|\mathbf{x}_n)\}$ as the 'model version' of interest. But, we could follow the same sort of analysis using, for example, the pair $\{f(\mathbf{x}_{n-s}|\theta), \pi(\theta|\mathbf{x}_s)\}$, where $\mathbf{x}_s=(x_1, \ldots, x_s)$, $\mathbf{x}_{n-s}=(x_{s+1}, \ldots, x_n)$, for, say, the smallest s such that $\pi(\theta|\mathbf{x}_s)$ is proper. This version is more concerned with 'fidelity to the data'. With logarithmic utility, the decision tree representation is

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms



The formal solution is to prefer M_1 if

$$\int \log\left\{\frac{q_1(\mathbf{x}_{n-s}|\mathbf{x}_s)}{q_2(\mathbf{x}_{n-s}|\mathbf{x}_s)}\right\} q_A(\mathbf{x}_{n-s}|\mathbf{x}_s)\,d\mathbf{x}_{n-s} > 0$$

and, based on K random selections from the $j=1, \ldots, \binom{n}{s}$ partitions $\mathbf{x}=[\mathbf{x}_s(j), \mathbf{x}_{n-s}(j)]$, a Monte Carlo approximation to the integral leads to the (approximated) solution: prefer M_1 if

$$\prod_{j=1}^{K} \{B_{12}(\mathbf{x}_{n-s}(j), \mathbf{x}_s(j))\}^{1/K} > 1.$$

This is now precisely the geometric mean version of the Berger and Pericchi (1993) intrinsic Bayes factor solution. See sections 6.1.6. and 6.3.3 of Bernardo and Smith (1994) for further details and discussion.

On previous occasions, I have heard a Professor O'Hagan berate audiences for their un-Bayesian sins. Were they clear about their fundamental concepts? Were their analyses focused on well-defined decision problems? Did their proposed solutions—with judicious approximation, if necessary—emerge from a coherent analysis?

In seconding the vote of thanks for this paper, I cannot help wonder what happened to *that* Professor O'Hagan.

The vote of thanks was passed by acclamation.

Aart F. de Vos (Free University, Amsterdam): O'Hagan's exposition of the problems that Bayesians have in deriving robust Bayes factors is of rare clarity. His fractional Bayes factors (FBFs) are a new trial, and a miracle of symmetry and simplicity. Unfortunately, they look like a trick. First, a solution is provided which is reasonably independent of prior distributions, while at the same time the importance of the prior for the Bayes factor is stressed. This is simply a contradiction. Second, 'fractional probability statements' are used, unclear things that did not occur before. This is a pity, especially for a Bayesian who cherishes coherence and reproaches others for the use of badly motivated tricks.

At the Bayesian riverboat conference in 1993 an earlier version of this paper confronted a similar paper of mine concerning the linear model (de Vos, 1993). O'Hagan was inspired to use his FBF for the linear model, I in my turn to try larger training sets than the minimal ones. The results, equation (20) and mine below, appeared to be strikingly similar.

The linear model is particularly interesting as different training sets have different information content. My formulae use weighted geometric means of real predictive probability statements based on all possible training samples of size m. This is the only possible way to use an extremely powerful theorem from linear algebra (Binet–Cauchy) such that the Bayes factor depends on the sufficient statistics S_i^2. In O'Hagan's notation ($b=m/n$), I obtained (without asymptotic arguments)

$$B_m(x) = c(Z_1, Z_2)\, g(n, m, r_1, r_2)(s_1^2/s_2^2)^{-(n-m)/2}$$

with $s_i^2 = S_i^2/(n-p_i+1)$ (the 'unbiased' estimate of σ_i^2, more appealing than S_i^2), $c \cong 1$ for well-behaved Z-matrices and $g(n, m, r_1, r_2)$ almost equal to O'Hagan's multiplier. However, plausible (though speculative) extensions of my arguments lead to alternative formulae based on large values of the training samples while retaining maximum discriminatory power between the models.

The good performance of the FBF and the similarity with my formulae suggest that we are nearing a spectacular answer to a question that is not yet clearly formulated. The key seems to be like entropy: the logarithms of predictive probability statements must in some way be averaged. The FBF does so

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

indirectly: it would be the simple geometric average of $q_i(x)/q_i(y_s)$ (from equation (7)) over all possible choices of the training sets y_s, if simple permutation arguments could be used. I thus expect that *the* answer will resemble O'Hagan's FBF, but no more than that.

D. V. Lindley (Minehead): I claim that what follows is a counter-example to fractional Bayes methodology. A scientist approaches a statistician with data $x \sim N(\theta, 1)$ and wishes to know whether $\theta<0$ or $\theta>0$. The (fractional) statistician explains that there are two models, M_1 that $\theta<0$, M_2 that $\theta>0$, and that it is required to calculate $P(M_1|x)$. The difficulties with uniform priors over θ are explained and the fractional solution is provided. The scientist is surprised by the complexity but is impressed, perceives the difficulty and accepts the statistical advice.

Somewhat later, the scientist reflects that the statistician had a distribution over θ, so why not calculate $\pi(\theta>0|x)$ and forget about the models? The scientist returns to the statistician who says that they now have an estimation problem. Even if $\pi(\theta)$ is improper, $\pi(\theta|x)$ is proper, the calculation is easy and the result differs from the earlier one.

Why is there this conflict between easy estimation and difficult model choice?

Consider, in the context of the scientist's problem, an improper prior $\pi(\theta)$. Write $\Pi(A)=\int_A \pi(\theta)\,d\theta$ whenever this exists. A probabilistic interpretation for $\pi(\theta)$ is that $P(A|B)=\Pi(AB)/\Pi(B)$ whenever $\Pi(B)$ exists. With a uniform prior, this probability exists for any bounded B. I contend that these conditional probabilities are *all* that π provides. In particular, the uniform prior does *not* provide a value for $P(\theta>0)$. Yet the fractional statistician, cited above, did provide such a value, calling it $P(M_2)$. Therein lies the contradiction. To see that $P(\theta>0)$ must remain undefined with θ uniform, let B be the set $-n<\theta<\rho n$ for some positive ρ. Then $P(\theta>0|B)=\rho/(1+\rho)$. If $n\to\infty$, so that B tends to the whole line, this tends to $P(\theta>0)$, which can assume any value in (0, 1), dependent on ρ. In other words, it is incoherent to assign both improper priors and a probability for the model. Almost all the papers referenced in this paper are similarly incoherent. Improper priors can be used but only with the greatest of care. It is better to think about θ and what it means to the scientist. It is his prior that is needed, not the statistician's. No one who does this has an improper distribution.

Steffen L. Lauritzen (Aalborg University): It is an important challenge to develop theoretically well-founded methods for model comparison. When comparing only a few models, the most important criteria are probably difficult to formalize and rest on connections between the model and the subject-matter context. But, for example, when analysing large, sparse contingency tables using graphical models (Whittaker, 1990) computers make it possible to compare astronomical numbers of models. A systematic approach then becomes mandatory.

What assistance could fractional Bayes factors provide here? To keep things simple we may look at the case when the model of independence M_2 is compared with the unrestricted model M_1 in an $I\times J$ contingency table with counts x_{ij}. If we use weak priors proportional to $\Pi_{ij}\,p_{ij}^{-1}$ in M_1 and proportional to $\Pi_{ij}\,p_{i+}^{-1}\,p_{+j}^{-1}$ in M_2 we obtain

$$B_b(x)=\frac{\Gamma(x_{++})\prod_{ij}\Gamma(x_{ij})}{\prod_i\Gamma(x_{i+})\prod_j\Gamma(x_{+j})}\,\frac{\prod_i\Gamma(bx_{i+})\prod_j\Gamma(bx_{+j})}{\Gamma(bx_{++})\prod_{ij}\Gamma(bx_{ij})}$$

provided that $x_{ij}>0$ for all i and j.

The first problem appears if we have a cell with zero counts (which is quite common for large sparse tables). We could use that $\Gamma(z)\sim z^{-1}$ for small values of z, and replace $B_b(x)$ by its limiting value when $x_{ij}\to 0$. This would give rise to a factor b^{n_0} where n_0 is the number of zero cells and the products should then only extend to cells with positive counts.

If we look around this difficulty, we must become very uneasy at the dependence of the Bayes factor on the arbitrary fraction b. Clearly, $B_1(x)=1$ but, when $b\to 0$, $B_b(x)$ is approximately proportional to $b^{(I-1)(J-1)}$ and so tends to 0. This gives great possibilities for manipulation.

The author suggests using $b=m_0/n$ where m_0 is the minimal sample size and n is the actual sample size. But sample size is a fragile concept and it is not quite clear what it means, even in this case. In a certain sense, any fixed sample is of size 1. One interpretation in this case is to let $m_0=IJ$ and $n=x_{++}$. But, if we consider the corresponding Poisson model with x_{++} being random, we have $m_0=n=1$ and therefore choose $b=1$ when following the author's recommendation. And the expression for $B_b(x)$ does not change.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

So without convincing guidelines for the choice of b that are not based on asymptopia, I remain sceptical and tend to think that I have seen yet another *ad hoc* statistical tool that throws little light on the important issue of model comparison.

David Draper (University of Bath): I would like to raise the issue of why we would wish to do model comparison in the first place, because this bears on *how* we should do it. As Professor O'Hagan notes, Bayes factors are central to the Bayesian approach to the comparison of a finite number of models, indexed discretely. Sometimes when the purpose of the underlying investigation is inference, the vector of implied posterior probabilities for the models, given a particular choice of prior probabilities, is sufficient to answer the scientific question at hand, for example when each model corresponds to a distinct substantive theory and the goal is to summarize the weight of evidence in favour of each theory. But often the role of model comparison is more technical, leading in routine current practice to a single specification choice such as the form of the error distribution in a generalized linear model. As noted in Draper (1995) and elsewhere, an alternative—arguably preferable in many cases—would be to deal with this sort of uncertainty more smoothly by indexing the models under consideration with one or more continuous parameters and adding a layer to the modelling hierarchy corresponding to the specification uncertainty. In effect one then computes an infinite number of Bayes factors, which give rise to mixing weights used in the calculation of a weighted average posterior distribution for the quantity of direct interest. I would be interested in any comments that Professor O'Hagan might wish to make on the extent to which his concerns with discrete Bayes factors, regarding the fact that the dependence on the prior distribution does not disappear with sample size, carry over to the continuous hierarchical case in which model uncertainty is dealt with more smoothly. I was surprised by these concerns even in the discrete case, given the $O(1)$ nature of the contribution to the Bayes factor of the prior distributions on the parameters specific to each model under comparison (see equation (5)).

A. P. Dawid (University College London): Suppose that we assign a single 'distribution', i.e. σ-finite measure, over the full parameter space, i.e. the disjoint union of model-specific parameter spaces. If this is improper then there is no sensible way of defining marginalization and conditioning, so we will not have well-defined prior model probabilities, or Bayes factors. However, even in this case the full posterior will typically be proper, so that we will have well-defined *posterior* model probabilities, thus answering our real need. The specification of such a prior still requires a 'ratio of constants', but at least we can now see why, which may give some guidance—unlike the situation considered by O'Hagan, with its inherent arbitrariness.

Using an improper prior or an arbitrary proper prior, as above, a 'principle of precise measurement' applies. The full posterior, given enough data, is insensitive to the prior: it concentrates on the appropriate model, with the usual asymptotic normal form. The effect of the prior is of additive order $O(1)$ on a log-posterior-model-odds, or log-Bayes-factor when it is defined, of order $O(n)$ or $O(-\log n)$. For extensive data, this is not worth worrying about. For small data sets it is surely appropriate that prior opinion is relevant and deserves careful elicitation. It seems a retrograde step to attempt to mask an $O(1)$ effect by introducing a fractional Bayes factor with unspecified b, variations in which will have an effect of at least the same order—or even much larger under suggestions (b) and (c) of Section 6.

Granted that the specification of priors is still an unfamiliar and delicate task, the following approach to coherence across models may be helpful. First, specify a proper prior for the most complex model considered, or for a new model generalizing all those considered. Then specify a proper prior within each of the other models to match, as closely as possible, the induced predictive distribution for a suitable 'minimal sample', varying with the model considered. This idea is related to, but distinct from, that underlying the partial Bayes factor, and is fully coherent. Essentially this approach was used by Dawid and Lauritzen (1994) for comparing different decomposable graphical models.

L. I. Pettit (Goldsmiths' College, London): I would like to congratulate Professor O'Hagan on an interesting paper that will be very useful.

All the examples considered concern global model choice. What happens if we consider local model choice? In particular I shall investigate whether a normal sample contains a single outlier. The asymptotics of Section 1.3 do not now apply.

Suppose, under M_0, $x_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and, under M_1, $x_i \sim N(\mu, \sigma^2)$ for $i \neq j$, $x_j \sim N(\mu + \delta, \sigma^2)$. For comparison with Spiegelhalter and Smith's (1982) method I shall take the prior proportional to $(\sigma^2)^{-(p_i+3)/2}$ which leads to r_i in equation (19) being constant. Note that this prior is necessary for

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

scale invariance in Spiegelhalter and Smith (1982), whose Bayes factor is

$$B_{01}^{SS} = \left\{\frac{3(n-1)}{2n}\right\}^{1/2} \left\{\frac{\sum_{1}^{n}(x_i-\bar{x})^2}{\sum_{i\neq j}(x_i-x^*)^2}\right\}^{-n/2} \tag{30}$$

where $x^* = (n-1)^{-1}\,\Sigma_{i\neq j}\,x_i$ (see Pettit (1992)) and O'Hagan's, taking $b = 3/n$, is

$$B_{01}^{OH} = \left\{\frac{\sum_{1}^{n}(x_i-\bar{x})^2}{\sum_{i\neq j}(x_i-x^*)^2}\right\}^{-(n-3)/2}. \tag{31}$$

I have argued elsewhere (Pettit, 1992) that this is a situation where we have some knowledge about the prior probabilities on the models (outliers are surprising) and that we should look for a log-Bayes-factor of -4 or less before we have good grounds for believing that an observation is an outlier.

To compare equations (30) and (31) I computed their values for a data set consisting of the expected normal order statistics for a sample of size 9 plus an extra observation y. Fig. 5(a) gives a graph of the log-Bayes-factors against y. For very small values of y equation (30) is the larger but this is reversed for bigger y-values. The value of y giving a log-Bayes-factor of about -4 is 3.2 for Spiegelhalter and Smith's and 4.1 for O'Hagan's method. The latter value is very conservative.

We can also compare regression models. Suppose, under M_0, $y_i \sim N(\alpha+\beta x_i,\ \sigma^2)$ for $i = 1, \ldots, n$ but, under M_1, $y_j \sim N(\alpha+\delta+\beta x_j,\ \sigma^2)$. The Spiegelhalter–Smith Bayes factor is

$$B_{01}^{SS} = \left\{\frac{4(n-1)\,\mathrm{CS}(x_{(j)},\ x_{(j)})}{3n\,\mathrm{CS}(x,\ x)}\right\}^{1/2} \left(\frac{s^2}{s^2_{(j)}}\right)^{-n/2} \tag{32}$$

where

$$s^2 = \mathrm{CS}(y,\ y) - \frac{\mathrm{CS}(x,\ y)^2}{\mathrm{CS}(x,\ x)},$$

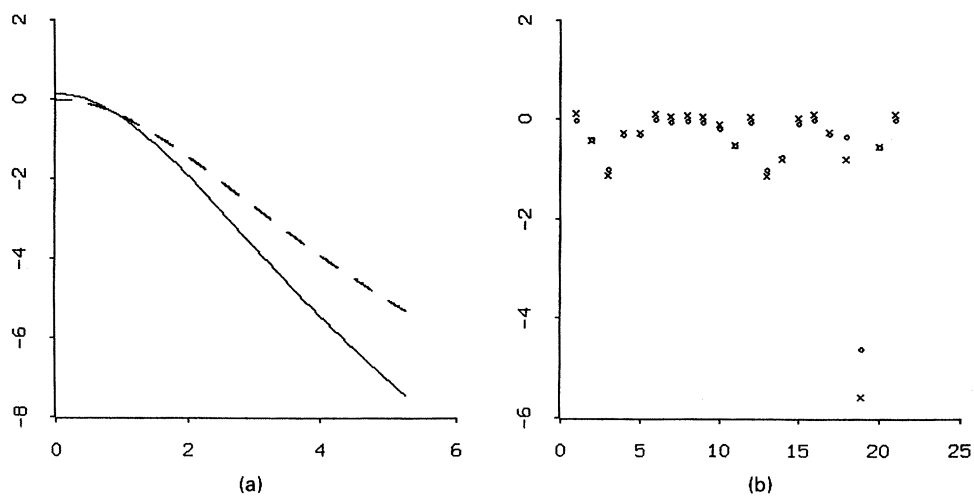


Fig. 5. (a) Plot of log B^{SS}(———) and log B^{OH} (— — —) for a model that y is an outlier *versus* y for a sample consisting of nine expected normal order statistics and y; (b) plot of log B^{SS} (×) and log B^{OH} (○) that observation i is an outlier for the data of Mickey *et al.* (1967)

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

$CS(u, v) = \Sigma (u_i - \bar{u})(v_i - \bar{v})$ and subscript (j) denotes deleting the jth observation (Pettit, 1992). O'Hagan's factor with $b = 4/n$ is

$$B_{01}^{OH} = \left(\frac{s^2}{s_{(j)}^2} \right)^{-(n-4)/2}. \tag{33}$$

Fig. 5(b) shows the values of equations (32) and (33) for the data of Mickey *et al.* (1967). Observation 19 shows up as an outlier although again O'Hagan's Bayes factor is more conservative. Observation 18 is the only other observation showing much difference; it has a large leverage and the term $CS(x_{(j)}, x_{(j)})/CS(x, x)$ is substantial.

To sum up in the examples that I have looked at (and also in Poisson samples; Pettit (1994)) the O'Hagan Bayes factor is more conservative than Spiegelhalter and Smith's, perhaps overly so. It does have the advantage that you do not need to think so much.

K. D. S. Young (University of Surrey, Guildford): Recently many different types of Bayes factor have been considered to circumvent the problem of improper priors. These include Spiegelhalter and Smith's (1982) B_{SS}, based on the device of imaginary observations, Geisser and Eddy's (1979) pseudo-Bayes factor B_{PSU}, based on averages of predictive densities, Berger and Pericchi's (1993) intrinsic Bayes factor B_{INT}, which uses minimal data sets to make the prior proper, and Aitkin's (1991) posterior Bayes factor B_{POST}. We must now add O'Hagan's fractional Bayes factor B_{FRAC}.

A comparison for these Bayes factors was made for the linear model

$$y \sim N(A_i \theta_i, \sigma^2 I_n)$$

with prior

$$p(\theta_i, \sigma | A_i) = c_i (2\pi\sigma^2)^{-(p_i - 1)/2} \sigma^{-1}$$

for $i = 0, 1$, where c_i are undefined constants. In the case of testing a specified mean we have found that in general

$$B_{INT} \geqslant B_{FRAC} \geqslant B_{SS} \geqslant B_{PSU} \geqslant B_{POST}.$$

A similar ordering is obtained for a simple linear regression testing whether the slope parameter is zero, the only difference being that B_{SS} and B_{FRAC} switch order when the data support the model with non-zero slope.

In Section 3.1 O'Hagan considers an example

$$M_0: x_1, \ldots, x_n \sim N(\theta_0, 1),$$
$$M_1: x_1, \ldots, x_n \sim N(\theta, 1)$$

where $p(\theta) \propto c$. A diagnostic measure to determine the effect of a proper prior on a Bayes factor is (Young and Pettit, 1993)

$$k_P = \ln B_{01}^P - \ln B_{01}^N$$

where B_{01}^P is a Bayes factor comparing models M_0 and M_1 based on proper priors and B_{01}^N is based on non-informative priors. We have calculated k_P using a proper prior for $\theta \sim N(\mu, \sigma^2)$ and using both B_{FRAC} with $b = 1/n$ and B_{SS} for B^N. The difference in k_P for the two Bayes factors is given by

$$k_P^{SS} - k_P^{OH} = \frac{1}{2}(\bar{x} - \theta_0)^2$$

which shows that the two will be different if $\bar{x}$ is far from θ_0, i.e. if M_0 is not true.

In Section 5.1 O'Hagan uses a different non-informative prior of the form $p(\theta) \propto \exp \theta$. In this case

$$k_P^{SS} - k_P^{OH} = \bar{x} - \theta_0 + \frac{1}{2n} + \frac{1}{2}(\bar{x} - \theta_0)^2.$$

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

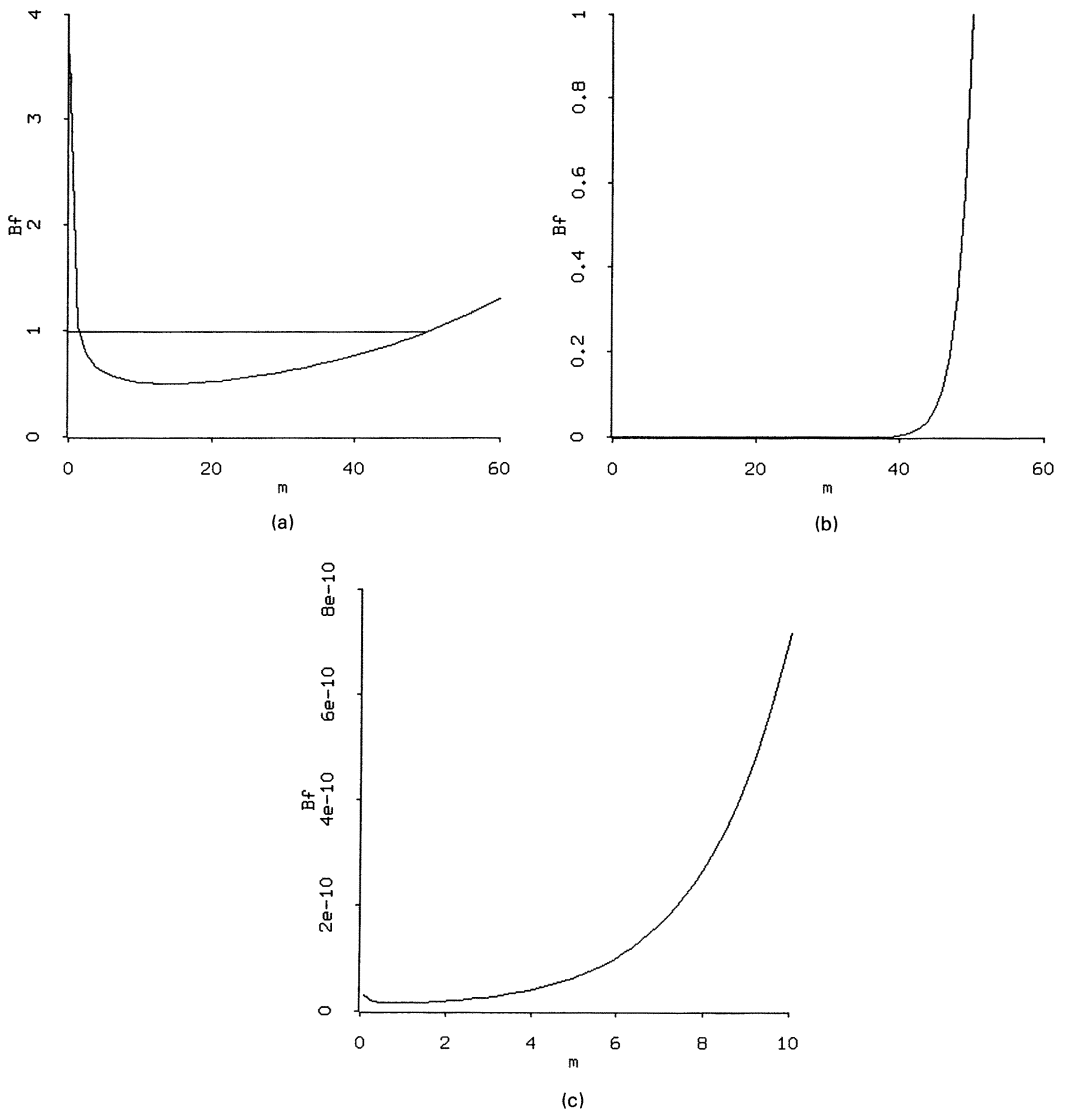


Fig. 6. Fractional Bayes factor for a simple linear regression set $y_i = \alpha + \beta x_i$ which is (a) in agreement with model M_0: $\beta = 0$, (b) in disagreement with model M_0: $\beta = 0$ and (c) in disagreement with model M_0: $\beta = 0$

As before the two will be different when $\bar{x}$ is different from θ_0, although because of the form of the prior the difference is no longer symmetrical.

The behaviour of the fractional Bayes factor as we vary b is important. In the simple linear regression model $E[y_i] = \alpha + \beta x_i$, we considered the case when the data are in agreement with model M_0: $\beta = 0$ and when there is disagreement. If we let $m \to n$ then $B_{\text{FRAC}} \to 1$ whatever, which is clearly seen in Figs 6(a) and 6(b). When there is agreement (Fig. 6(a)) it would seem that a small value of b is best. In Fig. 6(c) when there is disagreement there is a minimum when $b = 1/n$ and this again suggests that a small value of b should be chosen.

Some of this work has been done by Miss Julie Amiss as part of a doctoral thesis at the University of Surrey.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

L. R. Pericchi (Universidad Simon Bolivar, Caracas): Since Berger and Pericchi (1993) will not be published for some time and the author refers to it heavily, I shall describe some features of the intrinsic Bayes factor (IBF). The IBF solves the problem of a fully automatic Bayesian model selection, by taking averages (arithmetic or geometric) of minimal training samples. No prior assessments are required except default prior measures as in estimation. Furthermore, we give a theorem proving the existence of an intrinsic proper prior which gives results that are quite close to the arithmetic IBF, and we provide a formula to unveil an intrinsic prior, in quite general problems. The use of an intrinsic prior is completely insensitive to outlying training samples, obeys the sufficiency principle, is coherent and is even sequentially coherent, which O'Hagan's method is not. Take the example in Section 3.1, testing a normal mean θ_0 with variance 1. The intrinsic prior for θ turns out to be normal with mean θ_0 and variance 2. This is quite reasonable for the problem in hand, as in all examples that we have encountered so far. (If the variances were unknown, the intrinsic prior is a new distribution and obeys Jeffreys's *desiderata* for this problem.) For Darwin's data, the intrinsic prior produces a Bayes factor of 5.48 in favour of the simpler model, compared with only 3.814 by the author's method with his choice of $b = n^{-1}$. In a previous version of his paper (kindly provided by him) he strongly defended the simpler model for these data. All versions of the IBF, arithmetic or geometric, favour the author's favourite model more strongly than his method. Besides, the fractional Bayes factor is not automatic. The parameter b must be assessed, and in this simple example it is equivalent to the selection of the variance of a normal prior. So, what is the progress achieved by the present paper in the author's first example?

The examples in Sections 3.2 and 3.3 were not in the original paper, but they are analysed in detail in Berger and Pericchi (1993), where the IBF offers sensible automatic solutions.

There are deeper questions which are not addressed, e.g. when this method corresponds to a real prior, as in the normal example when $b \to 0$, or, when it is unavoidable to use resampling methods like the geometric IBF, as when none of the models entertained is assumed to be the sampling model, as I and A. F. M. Smith have suggested (work in progress).

Finally, even though this paper is far better than the previous version, the author has not yet justified his method besides interesting *ad hoc* arguments.

The following contributions were received in writing after the meeting.

Murray Aitkin (University of Western Australia, Nedlands, and Tel Aviv University): Conventional Bayes factors with diffuse priors are unworkable, and conjugate priors are not a workable substitute. The key to workable Bayes factors is the elimination of the arbitrary constant in the diffuse prior. This is achieved by integration with respect to the posterior (Aitkin, 1991) (Efron (1993) proposed the same idea under the name 'implied likelihood').

Professor O'Hagan achieves the same end by considering a fraction b of the data as a training sample, the remaining fraction $1 - b$ forming the likelihood. To avoid the obvious difficulty of deciding which subset of the data forms the training sample, he assumes the same data $\mathbf{x}$ in the full sample, the training subsample and the likelihood subsample. This leaves him open to the accusation of using the data thrice.

How well do fractional Bayes factors (FBFs) work? In Fig. 7(a), with 100 observations simulated from exponential(1), I have replicated Professor O'Hagan's Fig. 2 including the value of the posterior Bayes factor (PBF) with the same priors for comparison. The PBF is the full curve and the FBF the broken curve. The vertical scale is log-Bayes-factor.

The behaviours of the two Bayes factors are closely related, but the FBF is substantially better at identifying the exponential, a consequence of its heavy penalty on the two-parameter log-normal. In Fig. 7(b) the observations are generated from the log-normal distribution with (normal) $\mu = 0$ and $\sigma = 1$. Now the PBF is better for the same reason: the FBF has to overcome the heavy penalty on the correct model.

In conclusion, the FBF and the PBF have very similar properties except for comparing point null with general alternative hypotheses. Here the different penalties give characteristically different properties: the FBF can asymptotically confidently support a true null hypothesis, which the PBF cannot, but it pays the corresponding price of supporting less confidently a true alternative hypothesis. The relative merits of these properties may be debated, but it is puzzling to see the insistence, in this and other Bayesian discussions, that Bayes factors must have the properties of the Schwartz test criterion (5). This criterion is derived by using conventional priors and suffers all the disadvantages that Professor O'Hagan sets out in arguing for the FBF. Indeed, the constant a in approximation (5) is usually, as O'Hagan says, 'ignored', but it may have a substantial value compared with $\log n$ and is very sensitive to prior changes,

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

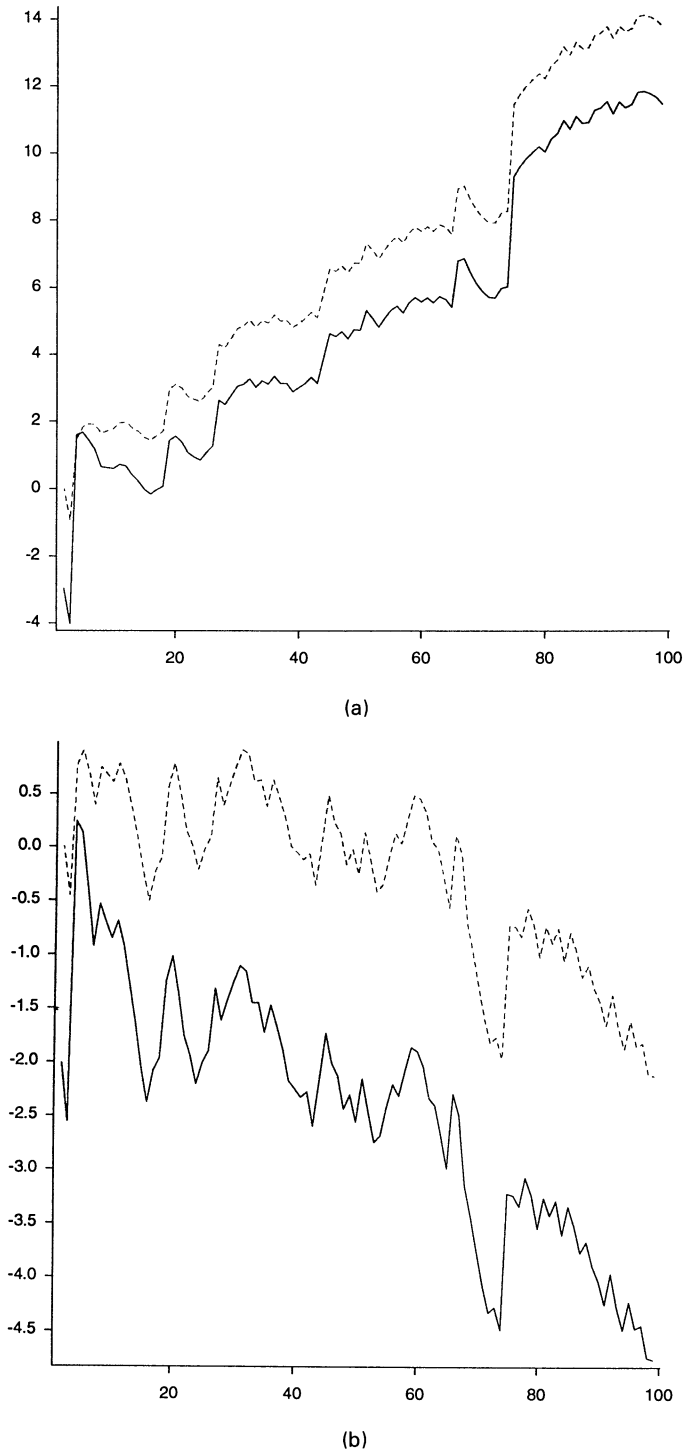


Fig. 7. Replication of O'Hagan's Fig. 2: exponential *versus* log-normal model

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

to say nothing of its unreasonable dependence on the scaling of variables in the information matrices of the two models. It is the removal of these information terms in the FBF in equation (20) that O'Hagan puts forward as a feature of the FBF, and I agree with him on its importance—see Aitkin (1993).

Other examples of applications of the PBF can be found in Aitkin (1992a, b) and Aitkin and Fuchs (1993).

G. A. Barnard (Colchester): During the 1914–18 World War, before anyone had suggested that significance tests and Bayesian reasoning were totally incompatible, Ethel M. Elderton collected data about industrial accidents showing that they followed a negative binomial distribution. This suggested the concept of 'accident proneness', according to which some of us are more prone to accidents than others are. But Elderton's data were equally compatible with our all being the same in this respect, if, following an accident, a further accident became more likely. The two models, so far as the available data were concerned, were indistinguishable.

On a strict interpretation the only one of Professor O'Hagan's examples that is free from difficulty in this respect is in Section 3.3. And even there the two models can be brought arbitrarily close to each other. One way of overcoming the difficulties would be to enquire what the client's prior (genuine, not 'conventional') for the parameters would be, on the supposition that model I is correct, and similarly for model II. Accepting each prior in turn converts each model into a simple statistical hypothesis and the likelihood ratio L given by the data for these two hypotheses would give a 'client's Bayes factor' which would seem to be more meaningful than that proposed by the author. If L turned out to be near to 1, it would suggest that we might be near an Elderton situation.

Some Bayesians have propounded the wholly unjustified dogma that the solution to every problem of statistical inference must take the form of a posterior distribution. In cases of the kind considered here we must allow that the appropriate response to the client's question may be that the available data do not allow a response of this form.

Referring to lines 7–9 of Section 7.2, ignoring part of the data does not necessarily lead to incoherence, any more than quoting a P-value does. But both these procedures may leave open the risk.

I am surprised that Professor O'Hagan makes no reference to the recent book by Geisser (1993) in which the model selection problem is treated at length and from a point of view which strikes me as more realistic than that taken here.

James Berger (Purdue University, West Lafayette) **and Julia Mortera** (Università di Roma III): This work is a potentially significant advance, although the author seems reluctant to embrace the real advance. The use of fractional Bayes factors with $b = m_0/n$ results in useful automatic Bayes factors. However, the other choices of b discussed are, at best, some type of *ad hoc* non-Bayesianism.

For testing a normal mean, consider the fractional Bayes factor (17). The identical expression would arise as an ordinary Bayes factor under an $N(\theta_0, \tau^2)$ prior, with $\tau^2 = (b^{-1} - 1)/n$. Jeffreys (and most Bayesians since) have recognized that the essential problem is simply that of making a sensible choice of τ^2. Jeffreys effectively recommended $\tau^2 = 1$, which is roughly equivalent to $b = m_0/n$. The alternatives discussed in the paper, for which $nb \to \infty$, correspond to letting $\tau^2 \to 0$. This is not sensible from a Bayesian perspective.

In this regard, the discussion in Section 5.1 is peculiar. Bayes factors have an inherent sensitivity to the prior, and this can be reduced only by inappropriately departing from Bayesian reasoning. We are reminded of the criticism one of us once received in an applied problem to the effect that the new statistical answer was unstable, having associated with it a non-negligible standard error, whereas before it had been their standard practice to use a non-statistical 'best guess', which was 'clearly better' because it had no instability. We are pleased to note that, in spite of the author's remonstrations on this issue, when it comes to actual examples he always uses the sensible $b = m_0/n$.

With $b = m_0/n$, the fractional Bayes factor appears to be similar to the intrinsic Bayes factors of Berger and Pericchi (1993). Arithmetic intrinsic Bayes factors have the very attractive property of always corresponding to actual (sensible) Bayes factors (at least for large n), and it is of interest to ask whether the same is true of fractional Bayes factors. For linear models with $t_i = 1$ the answer appears to be yes (details will be reported elsewhere), but for other choices of t_i the answer is no: the resulting fractional Bayes factors cannot arise as true Bayes factors (although they are not far off).

There are several references in the paper to issues such as 'sensitivity to outliers', 'sufficiency' and 'coherence' of the intrinsic Bayes factors and fractional Bayes factors. These are not serious issues (for

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

properly defined intrinsic Bayes factors) unless the sample size is quite small; for very small sample sizes Berger and Pericchi (1993) recommend the 'intrinsic prior' which overcomes all these difficulties.

David Cox (Nuffield College, Oxford): In the non-nested case, the Bayesian solution appears more incisive than that based on tests, but the latter may be more informative. By taking the two models in turn as the null hypothesis, we may study whether one, both or neither model is adequate. Clearly a model could have a large Bayes factor in its favour and yet be a very bad fit.

In the non-Bayesian analysis, which has a very extensive econometric literature, the asymptotic calculations (Cox, 1961, 1962) are best replaced by simulation.

D. V. Hinkley and I (Cox and Hinkley (1978), pages 160–162) put forward a very tentative Bayesian discussion leading to subtracting from the log-likelihood ratio a penalty $\log(n/n_0)\Delta d$, where Δd is the difference in the dimensionality of the parameters involved and n_0 is a notional sample size, said very boldly to be in the range $(\frac{1}{2}, 2)$, although $(\frac{1}{2}, 5)$ might have been better. The essence of the argument was that the prior probabilities in the two models should be the same over sets of parameters giving similar predictions. I hope that in his reply Professor O'Hagan will comment.

Andrew Gelman (University of California, Berkeley) **and Xiao-Li Meng** (University of Chicago): The idea of fractional Bayes factors (FBFs) is an intriguing attempt to avoid the fundamental problem of using Bayes factors with unspecified joint densities. However, the usefulness of the Bayes factor is restricted to problems where it exists. The non-existence of the Bayes factor, as is well known, is a direct consequence of not having a density (proper or improper) defined jointly for the model indicator and the parameters within each model. The proposed solutions of this problem, therefore, have either been to complete (temporarily) such a joint specification in some way, as with partial Bayes factors, or to define different quantities that no longer have proper probability (density) interpretations, as with the FBF. The FBF is well defined in its own right but no longer has a direct Bayesian interpretation, even under a properly specified joint density (except in the limit of $b=0$). When a method slides outside the Bayesian framework it is generally found that some incoherent aspects arise. The author discusses this issue in Section 8, but we are unsure whether sequential incoherence is the only drawback (for example, Section 7.2 does not convince us that the FBF is coherent for a given sample).

From an applied point of view, we do not see the necessity of working hard to define Bayes-factor-like quantities for models without joint densities. In our experience, a full Bayesian modelling approach can always address the questions of applied interest more directly than these look-alikes. If the models being compared are nested, then we prefer conducting Bayesian inference under the larger model, using a prior distribution with preference to the region of the parameter space near the smaller model, if appropriate; an elementary illustration is given in Gelman and Meng (1994). If the models under consideration are non-nested, it is generally reasonable to expand to a larger model class with an additional continuous parameter with specific values corresponding to the original models. For instance, for the data example in the paper, Darwin's data set, we prefer the approach of Box and Tiao (1962) using the power family, which includes wide-tailed distributions, to compute the posterior distribution of the parameters of applied interest.

Of course, in model comparison problems with proper joint densities (as, for example, in discrete models in genetics), we appreciate the utility of Bayes factors in posterior inference. We are also interested in seeing methods that can address applied interests, beyond what the full Bayesian modelling approach provides, in situations with no joint densities.

Rob Kass and Larry Wasserman (Carnegie Mellon University, Pittsburgh): In his examples Professor O'Hagan takes $b=n^{-1}$. This suggests that he may find it appropriate to take the amount of information in the prior to be about the same as that in one observation. Using this heuristic in a different way leads to an interesting result, at least for nested models where, say, $\theta_1=\beta$ and $\theta_2=(\beta, \psi)$ with the first model corresponding to H_0: $\psi=\psi_0$. We transform β so that the Fisher information matrix is block diagonal when $\psi=\psi_0$ (which is always possible) and take the marginal priors on β to be equal under the null and alternative hypotheses with β and ψ independent under the alternative. Then, taking the prior on ψ to be normal centred at ψ_0 and setting the determinant of the precision matrix equal to the determinant of the Fisher information matrix for ψ (so that 'the amount of information in the prior equals the amount of information in one observation') we find that the logarithm of the Bayes factor may be approximated by the Schwarz criterion with an error of order $O(n^{-1/2})$, rather than the usual error of order $O(1)$ (Kass and Wasserman, 1992). This result suggests that the Schwarz criterion should

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

provide sensible approximate solutions to Bayesian testing problems, at least for nested models. (Using a Cauchy prior, as suggested by Jeffreys, leads to the addition of a constant to the Schwarz criterion.)

A related approach is to use data-dependent priors. For example, define

$$\hat{\pi}_i(\theta_i) = h_i(\theta_i) f_i(x \mid \theta_i)^b \Big/ \int h_i(\theta_i) f_i(x \mid \theta_i)^b \, \mathrm{d}\theta_i.$$

The Bayes factor based on this data-dependent prior is equal to the partial Bayes factor up to relative order $1 + O(b)$ so the two are essentially the same. Such a prior is asymptotically coherent in that the dependence on the data vanishes for large n.

Sometimes we may want to restrict attention to a set of priors Γ. For example, suppose that the null hypothesis is $X_n \sim N(0, 1/n)$ and the alternative is $X_n \sim N(\theta, 1/n)$. Let $h_2(\theta) = 1$. We might take Γ to be all normals centred at 0. We then define $\hat{\pi}_2(\theta)$ to be the Kullback–Leibler projection of

$$h_2(\theta) f_2(x \mid \theta)^b \Big/ \int h_2(\theta) f_2(x \mid \theta)^b \, \mathrm{d}\theta$$

onto Γ. This yields the data-independent prior $\hat{\pi}_2(\theta) = N(0, 1)$. The Bayes factor based on this prior is again essentially the Schwarz criterion.

These results make us think that the Schwarz criterion is a good substitute for the Bayes factor in moderately large samples—which is the only situation in which an automatic method is appropriate. Furthermore, Professor O'Hagan's approach using $b = n^{-1}$ amounts to essentially the same criterion. We thus interpret the main result of his paper to be another justification for the Schwarz criterion. Perhaps he sees things differently.

Michael Lavine and Robert Wolpert (Duke University, Durham): O'Hagan undertakes the problem of finding a Bayes factor for comparing two models when at least one of the models uses an improper prior. Before asking whether his solution is sensible we want first to ask whether *any* solution is sensible.

Improper priors are often used in the hope that their posteriors approximate well the posterior that would have resulted from any well-thought-out proper prior. We typically reason that 'my prior is flat compared with the likelihood', 'there is much more information in the data than in the prior' and therefore 'my posterior is well approximated by the posterior from a convenient improper prior'. Then we adopt the improper prior and invest our effort more productively in other aspects of the analysis.

But approximation of the posterior is not the same as approximation of the Bayes factor. When O'Hagan considers an improper prior without saying which proper priors' posteriors he is hoping to approximate then we should bear in mind all proper priors with posteriors similar to that from the improper prior. If these priors were all to yield roughly similar Bayes factors, then it would be reasonable to associate a Bayes factor (or a small range of Bayes factors) with the improper prior, and it would be sensible to look for a convenient way to compute it. If, however, there are priors that yield posterior distributions similar to that from the improper prior, but that yield vastly different Bayes factors, then the specification of a Bayes factor for the improper prior is problematic at best.

The example in Section 3.1 illustrates the point. The problem is that of choosing either model 1, $x \sim N(0, 1)$, or model 2, $x \sim N(\theta, 1)$ for some uncertain $\theta \in \mathbf{R}$. O'Hagan's prior for θ is uniform on the real line, with density $h(\theta) \equiv 1$. Suppose that $x = 5$ is observed. For $m > 10$ all priors uniform on intervals $(-m, m)$ yield roughly the same posterior as his prior. Under model 1, the marginal density at the observation is $f(5) \approx 6 \times 10^{-12}$ whereas, under model 2, x is approximately uniformly distributed between $\pm(m+1)$, so $f(5)$ can range from around 0.05 to 0. The Bayes factor in favour of model 1 can range from around 10^{-10} to ∞.

We conclude that there is *no* uniquely acceptable Bayes factor, even approximately, although we admire O'Hagan's ingenuity in computing it! The hypothesis H_0: $\theta = 0$ will be preferred to H_1: $\theta \sim U[-10^{20}, 10^{20}]$, but not to H_2: $\theta \sim U[-10, 10]$. When comparing models the investigator simply cannot shirk the responsibility of specifying the prior distribution (and hence the alternative hypothesis) in more detail.

Adrian E. Raftery (University of Washington, Seattle): I congratulate Professor O'Hagan on an impressive paper and a clever idea. Overall, though, I am uneasy. The history of Bayesian estimation is marked by a cycle in which 'vague' improper priors are proposed, difficulties are found, adjustments are made, and the cycle starts again. The same cycle is apparent for testing. The device of Spiegelhalter

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

and Smith (1982) worked well in some difficult cases (Akman and Raftery, 1986; Raftery and Akman, 1986), but now we learn that it has problems.

Jeffreys (1961) himself did not recommend 'Jeffreys' priors for Bayes factors. Instead, he used proper priors that are fairly flat over the region where the likelihood could be substantial. I now feel that this approach is more promising than further efforts to trick improper priors into giving reasonable Bayes factors. It can be extended from the simple cases that Jeffreys considered to more complex models, such as linear regression (Raftery *et al.*, 1993). It is the basis for the GLIB software, which computes Bayes factors for generalized linear models (Raftery, 1993); this can be obtained by sending the message 'send glib from S' to statlib@stat.cmu.edu.

Although Bayes factors are indeed sensitive to the prior, this often does not invalidate conclusions. The Darwin data illustrate this. Transform these by $y^* = y/20 - 1$ and consider comparing model 1, $N(0, 1)$, with model 2, $N(\theta, 1)$, with an $N(0, \phi^2)$ prior for θ. Berger and Sellke (1987) used $\phi = 1$, which is close to the prior of Jeffreys (1961). I would not want to use $\phi > 3$ because this leads to 'substantial evidence' (Jeffreys's term) against model 2 when there is one observation at 0. I thus consider $\phi = 1$, 2, 3, for which $B = 3.9, 7.7, 11.5$. The ratio of largest to smallest B is less than 3 (equivalent to evidence 'not worth more than a bare mention') and the overall conclusion does not change much over this wide range of priors. The intrinsic Bayes factor of Berger and Perrichi (1993) is 20.8, which corresponds to $\phi = 5.4$. As the author said, this seems too big.

The Schwarz criterion yields a 'reference' Bayes factor, because it provides a good approximation when, roughly speaking, the information in the prior is equal to that in one 'typical' observation (Kass and Wasserman, 1992). Thus it corresponds to a reasonable proper prior and is easy to compute. It can also be viewed as a fractional Bayes factor with $b = n^{-1}$. For the Darwin data the fractional Bayes factor is 3.814, whereas the Schwarz approximation is 3.810.

Donald B. Rubin (Harvard University, Cambridge): Although interesting, Professor O'Hagan's development appears to suffer from the same basic limitation as many standard methods for the comparison of Bayesian models: it is predicated on the truth of one of the models being compared, and if all the models being compared are wrong then selecting the best according to Bayes factors can be practically disastrous. If we accept the commonly held view that all models are likely to be wrong yet some can be very useful for many purposes, the critical issue when comparing, selecting and accepting models is to be sensitive to the purposes to which they will be put. For this goal, some goodness-of-fit measure tuned to the specific intended purpose is needed to compare and select models.

For a specific example, in Rubin (1983) I drew inferences for a real finite population of 804 cities from a simple random sample of 100 by using Bayesian models with Box and Cox (1964) transformations to normality. The best fitting model according to straightforward likelihood or Bayes factors criteria gave atrocious real world inferences for the population total compared with the simple-minded, and obviously inferior fitting, normal model (see, in particular, my section 5). The use of fractional Bayes factors would not have helped here because, in a set of wrong models, the best model according to likelihood criteria can still produce predictions that are inconsistent with observed data or scientific understanding.

Model monitoring using posterior predictive check distributions (Rubin, 1984) is a more reliable way to compare models with respect to their success at intended purposes because it allows us to focus on specific quantities of interest rather than omnibus likelihood ratios or Bayes factors. With this technique, we compare the posterior predictive distributions of model monitoring quantities that are of practical relevance to see which of the posited models can produce adequate agreement with the observed data and scientific understanding. Although using likelihood ratio statistics (as in Rubin and Stern (1994)) or Bayes factors for the model monitoring quantities may often lead to choices similar to those made by using fractional Bayes factors, posterior predictive distributions of other quantities can lead to very different choices for the best model, or to the rejection of all models being contemplated—even if there is only one such model, or to the acceptance of a parsimonious model—even if there is strong evidence that it does not fit in unimportant ways. Recent work expanding and clarifying posterior predictive check distributions exhibits the generality of the idea for many problems (e.g. Meng (1994), Gelman and Meng (1994), Gelman *et al.* (1994) and Imbens and Rubin (1994)).

John W. Tukey (Princeton University): O'Hagan has presented us with an approach that seems better than previous Bayesian approaches. But I would very much like to see what more frequentist approaches

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

would do with the same data. How would plots of differences in log-tail-area for $|t|$ and χ^2 compare with the plots in Fig. 1?

A lesser point is raised by the same figure. Surely the author's evaluation of graph (c) has been biased by what he feels he ought to find. A ratio of e^{10}, which is what the trace attains, has to be interpreted, in my view, as '$N(2, 3)$ is more like $N(0.5)$ than $N(1, 1)$', although—not shown by the trace—the fit to both is very bad. Surely it is not a case of the data being 'unable to choose reliably' which model is 'more correct'.

The emphasis on such large values of $|\log B_b|$—on such near certainty of result—should not have a role in the real world. The pictures assume that the data come in piece by piece. Surely a well-regulated experimental programme would either stop collecting data, or, probably better, change the circumstances under which the data are being collected!

The **author** replied later, in writing, as follows.

I would like to thank the many discussants for their contributions. That so many eminent colleagues should have taken the trouble to contribute is a sign that this is a problem of genuine interest. Indeed, most agree that there is a problem (although not necessarily on what the problem is!), and I agree with them that it is a problem which probably does not have an ultimate solution. Fractional Bayes factors (FBFs) represent just one approach, which I believe to be the most useful so far within the framework of Bayes factors generally. Naturally, the discussants have some points of disagreement with me, and many make counter-claims for their own favourite methods. I shall try to deal as fully as possible with what has proved to be a fascinating and wide-ranging discussion.

Comparing models

I chose the words 'model comparison' in my title quite deliberately. I did not say 'model choice', 'model selection' or 'model testing'. Bayesian inference is a very flexible process and can provide various ways of choosing between models. It can also provide inferences in which a diversity of models are used throughout, as Dr Draper explains and expounds eloquently in Draper (1995). In a formal Bayesian analysis of any of these questions, Bayes factors enter either explicitly or implicitly, as one component alongside (proper) prior beliefs and utilities. The FBF is my attempt to provide a firm foundation for each of these inference problems in the presence of weak prior information.

I accept Professor Smith's comment that in a Bayesian analysis the prior distribution is just as much a part of the model as the likelihood. I have used the word much more in the classical statistician's sense of the way one structures beliefs about the data in terms of some unknown parameters. It would be nice to have a separate word for that.

Professor Rubin's posterior check distributions focus on the problem either of choosing a model or of testing the adequacy of a single model. Professor Cox also refers to the latter problem. Professor Rubin's idea has much in common with the classical approach of constructing *ad hoc* tests of fit. Such methods certainly have a role in exploratory work to identify models for 'comparison', but I believe that Bayes factors are central to subsequent formal inference.

Another important aspect of FBFs is that they are intended for the case where prior information is weak. Professor Barnard says that we should simply elicit the client's proper prior distributions and then there is no problem, but if the prior information is weak it is difficult to do this reliably. When conventional improper priors do not work, it does not help to force a proper prior distribution out of the client and to accept the result as a perfect representation of the client's prior knowledge. We must consider whether the inferences are robust to the substantial uncertainty which must attach to that prior specification. Professor Lavine and Professor Wolpert underline this fact. If in their example the client cannot assert which prior best represents his or her knowledge—uniform over a moderate range or uniform over a very large range—then the inferences will certainly not be robust.

Sensitivity and encompassing models

This brings me to my claim that the FBF gives greater robustness to prior specification. Both Professor Dawid and Dr Draper point out that the effect of the prior distribution in the Bayes factor is $O(1)$, and as n tends to ∞ this will always be swamped by the data. The Bayes factor will go either to 0 or to ∞ for all priors. This is true but not very helpful when in practice we tend usually to have a finite amount of data. Indeed, as Professor Tukey points out, we rarely collect so much data as to be able to make an absolutely conclusive choice between competing models. With $n < \infty$, $O(1)$ can be big. I still assert that the FBF can usefully increase robustness in moderate samples.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

Professor Dawid's argument also relies on the prior distributions for the various models being absolutely continuous with respect to an encompassing underlying measure. We must then specify their densities with respect to that measure, but that is equivalent to specifying the ratio c_1/c_2 in Section 1.1. He is therefore merely evading the original problem.

The related idea of embedding the various models in a grand encompassing model arises in the contributions of Professor Lindley, Dr Draper and Professor Gelman and Professor Meng. Any such solution must also imply finding a value for c_1/c_2 by some means. It is certainly true that the classical emphasis on hypothesis testing has led to too much consideration of nested models. The sharp null hypothesis should rarely be considered as a separate model with a non-zero probability mass. Professor Lindley's example is similar. Unless in the specific application there is some reason to consider that $\theta > 0$ and $\theta < 0$ are so intrinsically different as to induce potentially very different prior beliefs over the positive and negative half-lines, there is no reason to treat this as a model comparison problem. As in the nested case, both models fit into a wider framework in which they are defined by a continuous parameter that is meaningful, and so prior beliefs for it can be realistically elicited. However, Dr Draper and Professor Gelman and Professor Meng go too far when they imply that this can be done generally. Of course we can create an artificial grand encompassing model, but combining disparate models with quite different parameter spaces in this way does not help. We cannot use such an artificial structure to specify prior beliefs, and certainly not to give an unambiguous value to c_1/c_2. I agree with Draper (1995) that we must allow for model uncertainty by 'comparing' widely disparate models, which cannot be naturally subsumed in a continuous encompassing model framework.

Other methods

Several discussants advocate alternative methods, which surely all have their place in a statistician's toolkit. However, I still believe that the FBF is the most useful proposal so far in the class of methods based on partial Bayes factors and the idea of a training sample. Professor Smith, Dr Pericchi and Professor Berger and Dr Mortera favour intrinsic Bayes factors for various reasons. Dr Pericchi and Professor Berger and Dr Mortera regard the existence of an 'intrinsic prior' as important. One formulation of an arithmetic intrinsic Bayes factor is asymptotically equivalent to using a particular prior distribution, called the intrinsic prior by Berger and Pericchi (1993). Dr Pericchi advocates using this prior (assuming that it can be found) and computing the usual Bayes factor. This seems dangerous to me. The asymptotic equivalence to using a given prior is an interesting curiosity, but to suppose that this magically defines in any sense an ideal, natural or reasonable representation of one person's prior beliefs is to endow it with too much significance.

Professor Smith's characterization leads not to an arithmetic intrinsic Bayes factor but to the geometric form. Geometric averaging of Bayes factors is vastly more natural than arithmetic averaging, and this is the only form that I could be happy with. (This is another reason for distrusting the intrinsic prior.) Professor Smith's derivation is via one way of computing a solution to a specific decision problem, and I do not have a comparable derivation of the FBF to offer. As I said at the beginning of this reply, I see the FBF not as a solution to any specific inference problem but as underpinning a variety of inferences.

I am still suspicious of Professor Aitkin's posterior Bayes factor (see my discussion of conservatism below). Dr de Vos obtains other kinds of partial Bayes factor by using Binet–Cauchy methods. His work is certainly interesting but is rather a case of a technique looking for an application. His underlying rationale for averaging Bayes factors in the way that he does is just that he can derive a closed form solution in that case.

Professor Raftery's suggestion of using proper priors that are fairly flat over the region where the likelihood could be substantial is essentially what the FBF does. A prior proportional to the likelihood to the power b is certainly concentrated where the likelihood is substantial and is fairly flat if b is sufficiently small. The FBF does this automatically: perhaps in his (unpublished) references Professor Raftery achieves a similar effect. He also suggests that the Schwarz criterion can yield a 'reference' Bayes factor. Professor Cox's contribution is similar and can be seen as a way of writing the $O(1)$ term a in expression (5) by reference to his n_0, which is an effective sample size for some prior information. There may indeed be cases where there is real intuition about n_0, but in general its definition is rather abstract.

I must end this part of my reply in the same generous spirit as several of the discussants. The FBF may be a step forwards but it is far from an ultimate solution. There are many ways to look at these problems, and in practice a statistician might be well advised not to rely on just one.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

Conservatism

Dr Pettit and Dr Young both point out clearly the conservatism of the FBF, in that it gives factors which are generally closer to 1 than some competing methods. There are two reasons for this. One is shared by other forms of partial Bayes factor, which is that some of the data are being used for training. This conservatism may be adjusted for by the following argument.

Professor Lindley reasons convincingly that if the prior distributions on the $\boldsymbol{\theta}_i$s are improper then we cannot attach real meaning to the prior probabilities $P(M_i)$ of the models. The apparent implication is that even if a Bayes factor can be computed there is no prior odds $P(M_1)/P(M_2)$ with which to multiply it to obtain the posterior odds $P(M_1 \mid \mathbf{x})/P(M_2 \mid \mathbf{x})$. However, an important point is being missed. The partial Bayes factor $B(\mathbf{z} \mid \mathbf{y})$ should not be multiplied by $P(M_1)/P(M_2)$ to derive the posterior odds, but by $P(M_1 \mid \mathbf{y})/P(M_2 \mid \mathbf{y})$. Now the training sample $\mathbf{y}$ produces proper posterior distributions conditional on each model separately and Professor Lindley's comments no longer apply. The odds $P(M_1 \mid \mathbf{y})/P(M_2 \mid \mathbf{y})$ can be properly defined. The statistician or client might specify this ratio directly, reflecting beliefs about the two models based on the training data as well as prior knowledge. Multiplying it by the partial Bayes factor $B(\mathbf{z} \mid \mathbf{y})$ results in a fully coherent assessment of the posterior odds $P(M_1 \mid \mathbf{x})/P(M_2 \mid \mathbf{x})$. (Professor Aitkin's posterior Bayes factor cannot apparently be made coherent in this way.)

In practice it would not be easy to specify $P(M_1 \mid \mathbf{y})/P(M_2 \mid \mathbf{y})$. It would clearly depend on the choice of $\mathbf{y}$ and is in fact constrained so that every choice of $\mathbf{y}$ should lead to the same final posterior odds. Both the FBF and the intrinsic Bayes factors replace this ratio by some kind of typical value, which might in practice be easier or more difficult still to specify. An alternative is to defy Professor Lindley's logic and to insist on asserting a value for $P(M_1)/P(M_2)$, for instance by giving the models equal prior probabilities, and then to suppose that the training sample would have given similar evidence about the models, in proportion to its size, as the rest of the data. This is plausible for the FBF and would result in multiplying the asserted $P(M_1)/P(M_2)$ by $B_b(\mathbf{x})$ raised to the power $(1-b)^{-1}$. We might consider $B_b(\mathbf{x})^{1/(1-b)}$ as an FBF 'corrected' for the loss of the training information. (A similar 'correction' could be applied to intrinsic Bayes factors.)

The other source of conservatism in the FBF is the fact that it replaces a particular training sample $\mathbf{y}$ by a kind of typical training sample that produces the most perfect match to the remaining data. In a sense this gives each model the maximum possible 'benefit of the doubt'. Intrinsic Bayes factors incorporate variation in $\mathbf{y}$, which may be beneficial. The FBF could be likened to the simplistic classical approach to predictive distributions which replaces the parameter θ by its maximum likelihood estimate in the predictive likelihood, thereby failing to allow for posterior uncertainty about θ.

Other matters

I have dealt above with what I perceive to be the major themes arising in the discussion. It would take much longer to answer every single point, and so I shall now content myself with addressing just a few.

Professor Lauritzen makes some interesting observations concerning discrete data models. An even simpler example is a single binary sample, i.e. Bernoulli trials with probability θ of success. If the improper prior distribution proportional to $\theta^{-1}(1-\theta)^{-1}$ is used, what constitutes a minimal training sample? Two observations suffice to produce a proper posterior, but only if one is a success and one a failure. If this is the definition of a minimal sample, then all minimal training samples are alike, and all produce a uniform posterior. An alternative notion is that a sample is minimal if it has one success and at least one failure, or one failure and at least one success. Now the minimal training sample size is undefined. Provided that the full data contain a minimal sample (in either sense), the FBF can be defined for arbitrarily small b, but it is not clear what b corresponds to my 'minimal b' suggestion. Can any guidance now be given about b? As Professor Lauritzen found, letting b tend to 0 does not work at all. I think that the application of the FBF, and other kinds of partial Bayes factor, to discrete data problems will need separate and careful consideration.

In general the FBF is properly defined for b smaller than the 'minimal' size. Letting b go to 0 is equally unhelpful in general models, as Dr Young's Fig. 6(a) shows, for example. Professor Kass and Professor Wasserman point, however, to the interesting connection with their own work for a training sample equivalent to one observation, which is possible even if the minimal training sample would have more than one observation. They suggest that this still produces a sensible FBF. An open question, then, is how low can we go?

I am grateful to Dr Gilks for several very interesting comments. I particularly like his method for computing the FBF by the MCMC method.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

Finally, Professor Smith is dismayed (as others appear to have been) by my lapse into practices which he seems to suggest are not strictly within the Bayesian paradigm. I will not stoop to identifying similar heresies that he has committed, or (as I did jokingly at the meeting) plead for clemency on the grounds that it is my first offence! I do not think that I have strayed from clear Bayesian thinking. If I have, and if these methods do not strictly conform to the Bayesian paradigm, then I am confident that a sensible solution has not yet been found that does conform, either for the problem of model comparison with weak prior information or for the general question of Bayesian robustness.

REFERENCES IN THE DISCUSSION

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. R. Statist. Soc.* B, **53**, 111–142.

———(1992a) Evidence and the posterior Bayes factor. *Math. Scient.*, **17**, 15–25.

———(1992b) Model choice in contingency table analysis using the posterior Bayes factor. *Comput. Statist. Data Anal.*, **13**, 245–251.

———(1993) Posterior Bayes factor analysis of an exponential regression model. *Statist. Comput.*, **3**, 17–22.

Aitkin, M. and Fuchs, C. (1993) An analysis of models for the dilution and adulteration of fruit juice. *Statist. Comput.*, **3**, 89–99.

Akman, V. E. and Raftery, A. E. (1986) Bayes factors for non-homogeneous Poisson processes with vague prior information. *J. R. Statist. Soc.* B, **48**, 322–329.

Berger, J. O. and Pericchi, L. R. (1993) The intrinsic Bayes factor for model selection. *Technical Report 93-43C*. Department of Statistics, Purdue University, West Lafayette.

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

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

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Box, G. E. P. and Tiao, G. C. (1962) A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419–432.

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 105–123. Berkeley: University of California Press.

———(1962) Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B, **24**, 406–424.

Cox, D. R. and Hinkley, D. V. (1978) *Problems and Solutions in Theoretical Statistics*. London: Chapman and Hall.

Dawid, A. P. and Lauritzen, S. L. (1994) Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, **21**, 1272–1317.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

Efron, B. (1993) Bayes and likelihood calculations from confidence intervals. *Biometrika*, **80**, 3–26.

Geisser, S. (1993) *Predictive Inference: an Introduction*. London: Chapman and Hall.

Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection. *J. Am. Statist. Ass.*, **74**, 153–160.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Oxford University Press.

Gelman, A. and Meng, X.-L. (1994) Model checking and model improvement. In *Practical Markov Chain Monte Carlo* (eds W. Gilks, S. Richardson and D. Spiegelhalter). London: Chapman and Hall. To be published.

Gelman, A., Meng, X.-L. and Stern, H. S. (1994) Bayesian test for goodness of fit using tail area probabilities. To be published.

Imbens, G. W. and Rubin, D. B. (1994) Testing in randomized trials with imperfect compliance. To be published.

Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.

Kass, R. E. and Wasserman, L. (1992) A reference Bayesian test for nested hypotheses with large samples. *Technical Report 567*. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Meng, X.-L. (1994) Posterior predictive *P*-values. *Ann. Statist.*, **22**, in the press.

Mickey, M. R., Dunn, O. J. and Clark, V. (1967) Note on the use of stepwise regression in detecting outliers. *Comput. Biomed. Res.*, **1**, 105–111.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B, **56**, 3–48.

Pettit, L. I. (1992) Bayes factors for outlier models using the device of imaginary observations. *J. Am. Statist. Ass.*, **87**, 541–545.

———(1994) Bayesian approaches to the detection of outliers in Poisson samples. *Communs Statist. Theory Meth.*, **23**, no. 6, in the press.

Raftery, A. E. (1993) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Technical Report 255*. Department of Statistics, University of Washington, Seattle.

Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika*, **73**, 85–89.

Raftery, A. E., Madigan, D. M. and Hoeting, J. A. (1993) Model selection and accounting for model uncertainty in linear regression models. *Technical Report 262*. Deparment of Statistics, University of Washington, Seattle.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms

Rubin, D. B. (1983) A case study of the robustness of Bayesian methods of inference: estimating the total in a finite population using transformations to normality. In *Scientific Inference, Data Analysis and Robustness*, pp. 213–244. New York: Academic Press.

———(1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Rubin, D. B. and Stern, H. (1994) Testing in latent class models using a posterior predictive check distribution. In *Analysis of Latent Variables in Developmental Research* (eds A. Von Eye, C. C. Clogg *et al.*). To be published.

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

de Vos, A. F. (1993) A fair comparison between regression models of different dimensions. Submitted to *J. Econometr.*

Wasserman, L. (1992) Recent methodological advances in robust Bayesian inference. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 483–502. Oxford: Oxford University Press.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Young, K. D. S. and Pettit, L. I. (1993) On priors and Bayes factors. *Technical Report 93/10/St*. University of Surrey, Guildford.

This content downloaded from 147.8.204.164 on Sat, 29 Oct 2016 05:59:21 UTC
All use subject to http://about.jstor.org/terms