Bayes Factors in Practice

Robert E. Kass

The Statistician, Vol. 42, No. 5, Special Issue: Conference on Practical Bayesian Statistics, 1992 (2) (1993), 551-560.

Stable URL:
http://links.jstor.org/sici?sici=0039-0526%281993%2942%3A5%3C551%3ABFIP%3E2.0.CO%3B2-6

The Statistician is currently published by Royal Statistical Society.

The Statistician (1993) 42, pp. 551–560

Bayes factors in practice

ROBERT E. KASS

Department of Statistics, 232 Baker Hall, Carnegie-Mellon University, Pittsburgh, PA 15213-3890, USA

Abstract. Computational advances have facilitated application of various Bayesian methods, including the assessment of evidence in favor of a scientific theory. This is accomplished by calculating a Bayes factor. It is important to recognize, however, that the value of a Bayes factor may be sensitive to the choice of priors on parameters appearing in the competing models. This sensitivity is illustrated and discussed. The Schwarz criterion remains a crude yet useful approximation, but more accurate methods will generally require determination of priors and subjective sensitivity analysis.

1 Introduction

Recent advances in computation have made Bayesian methods practical in a wide variety of settings. In particular, it is now possible to take a Bayesian approach to testing sharp hypotheses and to compute Bayes factors in non-trivial and multi-dimensional problems. On the one hand, this opens up exciting possibilities for scientific inference. On the other, it calls for greater scrutiny of the use of Bayes factors. In this paper, I will draw on several examples to discuss the status of Bayes factors as data-analytic inferential tools. I will not go into foundational issues in any depth, nor will I attempt a thorough review of the subject. (For a review emphasizing the contrast between p-values and Bayes factors, see Berger and Delampady, 1987; for a practically oriented review see Kass and Raftery, 1993; important references neglected here may be found in these papers.) Instead, I will provide a perspective on the use of Bayes factors that is intended to serve as something of a guide and warning for anyone who has not yet thought carefully about them. The major issue I will emphasize is the sensitivity of Bayes factors to the choice of priors on parameters appearing in the two competing models.

2 The role of testing in Bayesian methodology

Bayes factors are commonly discussed as alternatives to p-values, which are criticized as being (i) poor measures of strength of evidence against a null hypothesis and (ii) improperly applied to problems that should be treated as involving estimation rather than testing. I will not dwell at all on (i) but will, below, make some remarks about the distinction between estimation and testing, because it is quite important and helps illuminate the purpose of calculating a Bayes factor. Here, I follow Jeffreys (1961) by taking the position that Bayesian hypothesis tests provide quantitative evaluation of the evidence in favor of a scientific law or theory. An example should help to make this point of view clearer.

Example: E. coli mutagenesis. In an experiment in molecular biology (Sklar & Strauss, 1980) the investigators hypothesized that for particular traits in certain strains of *E. coli*, mutations would occur by an unusual error-prone DNA repair mechanism; this would lead to an absence of linkage of mutations at neighboring loci. They created a pair of cell lines, one of which contained cells 'selected' for the relatively rare trait of rifampin resistance, the other of which contained 'unselected' cells. The absence of linkage,

predicted by the DNA repair hypothesis, would imply that proportions p_1 and p_2 of bacteria exhibiting acetate utilization deficiency in the 'selected' and 'unselected' cell lines would be equal, i.e. $H_0 : p_1 = p_2$. When the investigators took samples from each cell line and found $\hat{p}_1$ and $\hat{p}_2$ to be approximately equal they believed this ought to have represented fairly strong evidence in favor of DNA repair.

This example illustrates an important situation, one where a scientific hypothesis may be translated directly into a statistical hypothesis and the problem is to quantify the evidence in favor of (or against) the hypothesis. In particular, the hypothesis in question may be, as it is in this example, the *null* hypothesis, in which case specific non-Bayesian methodology is clearly lacking.

There is more to this example, which illustrates an additional important use of Bayesian methods generally, that of incorporating additional information. It turned out that data from many other strains were available, which were used (in calculations not reported in the paper cited) to formulate a prior under the alternative hypothesis. Bayes factor calculations led to the conclusion that the authors were quite justified in feeling that they had substantial statistical evidence in favor of their hypothesis. In Section 7, I will discuss a different example in which external information was incorporated into the analysis. Now, I wish to continue with the discussion of the circumstances in which Bayes factors are useful by contrasting an example in which it is *not* appropriate to phrase the problem as one of testing a sharp null hypothesis, i.e. one having the form $H_0 : \psi = \psi_0$ for some parameter ψ.

Example: ECMO. A much-discussed randomized clinical trial of extra-corporeal membrane oxygenation (ECMO) as a treatment for severe lung disease among newborns was conducted by colleagues of Ware (1989). After the first stage in the trial, all nine babies treated with ECMO survived, while only six out of ten treated with conventional therapy survived. At this point, the question was raised whether there was sufficient evidence in favor of ECMO to stop randomly allocating babies to the two treatments.

This example, like many clinical trials, differs from the *E. coli* example, in that it is not of fundamental interest whether the two treatments produce essentially the same survival probabilities. The treatments are, in fact, very dissimilar, and it is reasonable to assume the survival probabilities are at least somewhat different. The issue in the trial is *how much* difference there is in survival, and for *which treatment* is the survival probability larger.

In the ECMO example, it is conceptually appealing to consider instead the 'one-sided' hypothesis that the difference of the survival probabilities (or the difference of the logits) is larger than some specific value, or to estimate the size of the difference. What I would like to emphasize here is that this is quite different from testing the sharp hypothesis of *equality* of probabilities. In Jeffreys' terminology, the problem of computing the interval probability is one of 'estimation'. The term 'testing' is reserved for testing sharp hypotheses which require priors on lower-dimensional spaces. I do not mean to imply, however, that *all* two-binomial clinical trials should be treated as estimation problems in this sense. Occasionally, genuinely interesting sharp null hypotheses arise. For instance, if a trial were conducted to compare vitamin C to placebo in its ability to prevent subjects from becoming infected with a virus, the null hypothesis of no difference—which would correspond to there being, for practical purposes, no relevant physiological effect of vitamin C—would, at least in my opinion, be quite plausible. A numerical example in this setting is given by Kass and Vaidyanathan (1992).

This point deserves emphasis because, unlike in the frequentist approach, where tests of the form $H_0 : \psi = \psi_0$ generate confidence sets and confidence sets generate tests, there is no complementarity between Bayesian testing of sharp hypotheses and estimation methods. When a sharp hypothesis is involved, the prior must put mass on that hypothesis, e.g. $\psi = \psi_0$, whereas for 'estimating' ψ a continuous prior is used. Thus, in

each situation one must decide which analysis is the more appropriate. Furthermore, as illustrated in Section 5, sensitivity to choice of prior is a bigger concern in testing than in estimation, and thus the priors used in carrying out a test must be chosen with some care.

3 The general form and interpretation of Bayes factors

In general, the schematic form of the Bayes factor for comparing hypotheses H_1 and H_2 is

$$\text{Bayes factor} = \frac{P(data \mid H_1)}{P(data \mid H_2)}$$

and when the probability of H_1 (given H_1 or H_2 holds) is transformed to odds we have

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

When there are unknown parameters in the specifications of the data distributions, the formula for the Bayes factor becomes

$$B = \frac{\int p_1(y \mid \beta)\, \pi_1(\beta)\, \mathrm{d}\beta}{\int p_2(y \mid \theta)\, \pi_2(\theta)\, \mathrm{d}\theta} \tag{1}$$

where β and θ are parameters of the probability densities $p_1(y \mid \beta)$ and $p_2(y \mid \theta)$ for the data y that hold under the two respective hypotheses, and $\pi_1(\beta)$ and $\pi_2(\theta)$ are prior probability densities introduced to reflect initial uncertainty about the values of these parameters. In many problems the two sets of data distributions, parameters and priors that appear in the numerator and denominator of (1) are related in a simple way. Thus, when $\theta = (\beta, \psi)$ and under H_1 we have $\psi = \psi_0$, $\pi_1(\beta)$ is often taken to be the marginal density found by integrating $\pi_2(\beta, \psi)$ over ψ. In general, though, $\pi_1(\beta)$ and $\pi_2(\theta)$ need not be related and the form of (1) is thus quite general.

Jeffreys (1961, Appendix B) recommended interpreting the Bayes factor in units of $\frac{1}{2}$ on the $\log_{10}$ scale. Thus, he said that $-\log_{10} B > \frac{1}{2}$ represented 'substantial' evidence against the null hypothesis (here, against H_1 and in favor of H_2), $-\log_{10} B > 1$ was 'strong' evidence and $-\log_{10} B > 2$ was 'decisive'. These rough categories seem to furnish appropriate guidelines. It is perhaps worth mentioning that probability itself provides a meaningful scale made operational through betting and, thus, the labelling of these categories is not a calibration of the Bayes factor but rather a rough descriptive statement about standards of evidence in scientific investigation.

4 Computation

The integrals in (1) often must be evaluated numerically. In this case, several approaches are possible. First, it is usually not difficult to maximize the likelihoods appearing in the integrands of (1) and thus the Schwarz criterion (sometimes called the Bayes information criterion, or BIC) may be calculated. As discussed in Section 5.2, this furnishes a rough approximation to the Bayes factor that may be sufficient for many applications. Next, again assuming maximization may be carried out, Laplace's method (see equation (3)) gives a more accurate approximation. It is accurate enough for most practical purposes. Some discussion is given in Kass and Vaidyanathan (1992), and an application is briefly reviewed here in Section 7.

Generally, it is advisable to at least check asymptotic approximations with a method that does not directly depend, in theory, on having large sample sizes. One such is numerical quadrature, of which there are different varieties (Naylor & Smith, 1982; Dellaportas & Wright, 1991; Genz & Kass, 1993). The approach discussed by Genz and

Kass (1991) is based on widely available numerical integration software that has been modified for statistical applications. These authors call the method 'subregion-adaptive integration'. It is currently undergoing further development, and appears quite promising for reasonably well-behaved integration problems of less than about 15 dimensions. An alternative is to use Monte Carlo importance sampling (e.g. Geweke, 1989). This can be quite effective in many problems, as well.

Finally, much attention has been given lately to posterior simulation by Markov chain Monte Carlo. Unfortunately, there is, to my knowledge, no straightforward and effective general method of computing Bayes factors using samples from the posterior distributions under the competing hypotheses. If the likelihood function is easily evaluated and the posterior is not drastically far from normal, then an improvement on Laplace's method is available (Kass & Wasserman, 1992a). Perhaps preferable is the use of the formula

$$\int p(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta = \left(E\left(\frac{f(\theta)}{p(y|\theta)\,\pi(\theta)}\right)\right)^{-1}$$

where the expectation is with respect to the posterior and $f(\theta)$ is any probability distribution. The expectation is approximated by a sum, using draws of θ from the posterior. A simple choice for $f(\theta)$ is the modal normal approximation to the posterior. This generalizes an idea discussed by Newton and Raftery (1991) and was mentioned in Gelfand and Dey (1993). However, no thorough study of the method has been published. See Kass and Raftery (1993) for further references.

5 The problem of sensitivity to choices of priors

As a general rule, one must be cautious in choosing the priors π_1 and π_2 in (1) for they may have a more substantial influence on the results than they would if the problem were treated instead as one of estimation. I next illustrate this situation in the ECMO example mentioned earlier and then, in Section 5.1, point out that one would expect similar behavior to occur frequently. In Sections 5.2 and 5.3, I present methods and results that are helpful in nested models. In Section 6, I will note that the Schwarz criterion furnishes a very general, though very rough, approximation to the Bayes factor which avoids the introduction of the priors in equation (1), and in Section 7, I will briefly discuss an example in which a fully subjectivist approach was used together with substantial sensitivity analysis.

Example: ECMO (continued). Kass and Greenhouse (1989) reanalyzed the ECMO data using independent priors on the difference δ and the average γ of the log-odds survival for the two treatments. ECMO would be the better treatment if $\delta > 0$ and it was considered to be substantially better if $\delta > 0{\cdot}4$. Table 1 shows results obtained for four choices of priors (discussed by Kass and Greenhouse). Also shown is the Bayes factor in favor of equal survival probabilities for the two therapies.

One may observe that the results vary more for the Bayes factor than for the 'estimation' interval probabilities. Indeed, for the last prior, the Bayes factor turns *in favor*

Table 1. Posterior probabilities and Bayes factors for ECMO example (from Kass & Greenhouse, 1989)

Prior	$P\{\delta > 0 \mid y\}$	$P\{\delta > 0{\cdot}4 \mid y\}$	Bayes factor
B	0·97	0·93	$(3{\cdot}7)^{-1}$
C	0·95	0·90	$(2{\cdot}5)^{-1}$
D	0·94	0·88	$(2{\cdot}1)^{-1}$
F	0·96	0·93	2·1

of the null hypothetical equality of the two probabilities ($H_0 : \delta = 0$). In that case, the interval probabilities indicate ECMO is to be favored. The seeming paradox is explained by recalling that the Bayes factor is a ratio of the marginal probability of the data y under H_0 to its probability under the alternative H_A. For this prior, the observed data were unlikely under H_0 but they are *even more unlikely* under H_A. Thus, the ratio is greater than 1. (This may be regarded as a purely Bayesian version of the Jeffreys–Lindley paradox (Jeffreys, 1961, Appendix B; Lindley, 1957).)

5.1 Understanding sensitivity

A simple asymptotic analysis shows that even in large samples Bayes factors remain sensitive to the choice of prior. An approximation to the marginal density of the data under a hypothesis H having parameter θ

$$p(y|H) = \int p(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta \tag{2}$$

may be obtained by substituting for the integrand an approximation to it based on the usual modal normal approximation to the posterior. This posterior density approximation is the normal$(\tilde{\theta}, \tilde{\Sigma})$ density where $\tilde{\theta}$ is the mode and $\tilde{\Sigma}$ is the inverse negative Hessian (matrix of second derivatives) of the log posterior density $\log p(\theta|y) \propto L(\theta)\,\pi(\theta)$ evaluated at the mode. This method is called Laplace's method (Tierney & Kadane, 1986; Tierney *et al.*, 1989) and the resulting approximation is

$$p(y|H) \doteq (2\pi)^{m/2} |\tilde{\Sigma}|^{1/2} L(\tilde{\theta})\,\pi(\tilde{\theta}) \tag{3}$$

where m is the dimension of θ.

The important observation about this approximation is that the prior appears in it, in contrast to the analogous approximation for the posterior mean of a real-valued function $g(\theta)$ (which is also derived by Laplace's method),

$$E(g(\theta)|y) = g(\tilde{\theta})$$

Both of these approximations have errors of order $O(n^{-1})$. Furthermore, lest one think that, because the mode $\tilde{\theta}$ involves the prior, the prior is exerting substantial influence on the approximate posterior mean, it should also be noted that both approximations hold if the MLE is substituted for $\tilde{\theta}$. Thus, for large samples, we obtain the familiar insensitivity of the posterior mean to the choice of prior, whereas there is no directly analogous result in the case of Bayes factors: to compute the Bayes factor correctly in large samples one needs to evaluate the prior.

I will return to asymptotics in Section 6 and we will find that if one is satisfied with an order-of-magnitude approximation to the log of the Bayes factor, the prior may be ignored for large samples. However, the elementary asymptotic statement above underscores the warning served by the example in the previous section.

5.2 Nested models

In many problems the parameter appearing in the denominator of (1) may be written $\theta = (\beta, \psi)$ and the data density $p_1(y|\beta)$ is an instance of the density $p_2(y|\beta, \psi)$ when ψ takes some value ψ_0, so that the numerator is specified as $H_0 : \psi = \psi_0$. Taking $\pi_\beta(\beta)$ to be the prior under H_0, a simplification is to assume that β and ψ are *a priori* independent under the alternative H_A with prior density

$$\pi(\beta, \psi) = \pi_\beta(\beta)\,\pi_\psi(\psi) \tag{4}$$

That is, the marginal prior on β under H_A is equal to the prior on β under H_0. Jeffreys treated a variety of problems having this form in the case in which ψ is one-dimensional,

including that of assessing extra-Poisson variability (Jeffreys, 1961, p. 319). A similar problem is the following.

Example: Extra-binomial variability. When ostensibly binomial data are thought to be potentially over-dispersed, one may consider as a simple alternative the beta-binomial model. Here, under the binomial model, $Y_i \sim B(n_i, p)$, i.i.d. for $i = 1, \ldots, k$. Under the alternative, $Y_i \sim B(n_i, p_i)$ independently for $i = 1, \ldots, k$, and $p_i \sim \text{Beta}(\alpha, \beta)$, i.i.d. The strength of evidence against (or in favor of) the binomial may be assessed using Bayes factors. Consider reparameterizing the beta distribution according to $(\alpha, \beta) = (\xi/\omega, (1-\xi)/\omega)$. In this parameterization, under the beta-binomial, $E(Y_i/n_i) = \xi$ and the binomial becomes a limiting case of the beta-binomial as $\omega \to 0$. In the notation used at the beginning of this subsection, ψ becomes ω and β becomes ξ, which reduces to p when $\omega = 0$. The additional simplification is then to take ξ and ω to be *a priori* independent, with the distribution on ξ being the same as that used on p under the binomial. One may put a *Uniform*$(0, 1)$ prior on ξ and p and then examine the Bayes factor as a function of ω by plotting B^{-1} vs ω.

In cases such as this where there is a one-dimensional parameter of interest ψ (here ω) that takes a specified value ψ_0 under H_0, it is convenient to examine the value of the Bayes factor as a function of the values taken by ψ under the alternative H_A, since this avoids direct specification of the prior on ψ under H_A. Furthermore, once several such values of the Bayes factor are computed, these may be weighted to produce the Bayes factor for any specified prior on ψ.

The uniform prior on p and ξ is a plausibly interesting choice, but the discussion of the previous section suggests its effect must be considered carefully. Nonetheless, as Kass and Hsiao show (in unpublished work), in this example the Bayes factor turns out not to be very sensitive to the choice of prior on ξ. The reason is related to results discussed by Kass and Vaidyanathan (1992) indicating that, for β and ψ as in equation (4), under certain conditions, the Bayes factor is sensitive only to choice of prior on the parameter of interest ψ and not to the choice of prior on β. In particular, this insensitivity is likely to occur when β and ψ are 'null-orthogonal', which means that the Fisher information (expected information) matrix is block diagonal when $\psi = \psi_0$. As a practical matter in analyzing data, one may examine the modal covariance matrix under H_A. If β and ψ are approximately uncorrelated, they may be called *observed-orthogonal* (since then the observed information matrix is roughly block diagonal) and in this case, following the argument of Kass and Vaidyanathan, the Bayes factor will not be very sensitive to the choice of prior on β. This may greatly reduce the effort needed in checking sensitivity to choice of priors.

6 The Schwarz criterion

It is possible to avoid the introduction of the prior densities $\pi_1(\beta)$ and $\pi_2(\theta)$ in (1) by using instead of $\log B$ the quantity

$$S = \log p_1(y|\hat{\beta}) - \log p_2(y|\hat{\theta}) - \tfrac{1}{2}(m_1 - m_2)\log(n)$$

where $\hat{\beta}$ is the maximum of $\log p_1(y|\beta)$ as a function of β, $\hat{\theta}$ is the maximum of $\log p_1(y|\theta)$ as a function of θ, m_1 is the dimension of θ, m_2 is the dimension of γ, and n is the sample size. As $n \to \infty$, this quantity, often called the Schwarz criterion, satisfies

$$\frac{S - \log B}{\log B} \to 0 \tag{5}$$

and thus may be viewed as a rough approximation to the logarithm of the Bayes factor.

Equation (5) shows that in large samples the Schwarz criterion should provide a reasonable indication of the evidence. It is appealing insofar as it could be applied as a standard procedure even when precise formulation of $\pi_1(\beta)$ and $\pi_2(\theta)$ is difficult.

Schwarz (1978) derived (5) rigorously for linear subfamilies of exponential families. Since (5) is quite important, I would like to give a very simple sketch of the argument based on approximation (3). There are four steps.

(1) First, we apply (3) to both the numerator and denominator of (1). For simplicity, let us consider the denominator.

(2) Assume $n \cdot \tilde{\Sigma} \to \Sigma$ for some Σ. This regularity condition will hold fairly generally; for instance, in i.i.d. sampling Σ^{-1} will be the Fisher information matrix at the true value of θ. We then have $|\tilde{\Sigma}| \approx |\frac{1}{n}\Sigma| = n^{-m} \cdot |\Sigma|$. This will explain the term $\frac{1}{2}(m_1 - m_2)\log(n)$.

(3) If we take logs and ignore terms of constant order (i.e. terms that are not becoming positively or negatively infinite as $n \to \infty$) we obtain

$$\log p(y|H) \approx \log L(\tilde{\theta}) - \frac{m}{2} \log n$$

(4) Finally, the posterior probability of a hypothesis is a consistent indicator of truth of the hypothesis in the sense that as $n \to \infty$, $\log B \to \infty$ under H_1 while $\log B \to -\infty$ under H_2. Thus, if we apply step (3) to $\log B = \log p(y|H_1) - \log p(y|H_2)$, ignoring constant-order terms but retaining all terms that become positively or negatively infinite, we obtain (5).

From the outline given here, it is apparent that (5) holds much more generally than in the restricted setting treated by Schwarz. The argument shows that the Schwarz criterion is a rough approximation to $\log B$ that remains a consistent indicator of truth of the hypothesis as $n \to \infty$. I should also mention that under certain conditions on nested models as described in Section 5.2, if the prior on ψ is normal and the amount of information in that prior is equal to the amount of information in one observation from the sample, then the Schwarz becomes a more accurate approximation to $\log B$, with error $O(n^{-1/2})$ (see Kass & Wasserman, 1992b).

7 Subjective sensitivity analysis

When there is doubt about whether the sample size is adequate to justify reliance on the Schwarz criterion, or when greater accuracy is desired for any reason, the priors $\pi_1(\beta)$ and $\pi_2(\theta)$ must be determined. Since this is rarely accomplished with assurance of precision, it is necessary to carry out some kind of sensitivity analysis to examine the way results would vary if alternative priors were used.

Subjective determination of priors, especially in higher dimensions, is difficult. Once it is accomplished, however, the additional work of performing sensitivity analysis can be done fairly easily with the aid of asymptotic approximations. These are discussed in Kass and Vaidyanathan (1992). Briefly, if we wish to assess the effect of using priors $\pi_{1,NEW}(\beta)$ and $\pi_{2,NEW}(\theta)$ in place of $\pi_1(\beta)$ and $\pi_2(\theta)$, we may compute the resulting new Bayes factor using

$$B_{NEW} \approx B \cdot \frac{r_1(\tilde{\beta})}{r_2(\tilde{\theta})}$$

where $r_i = \pi_{i,NEW}/\pi_i$, for $i = 1, 2$, and $\tilde{\beta}$ and $\tilde{\theta}$ are the posterior modes under each hypothesis. This approximation has a multiplicative error of order $O(n^{-1})$, which is quite adequate for the purpose of checking for large discrepancies. The approach was illustrated in an application in the field of human–computer interaction by Carlin *et al.*

(1992), which I now outline. This example shows that all of the forgoing is indeed computationally effective in a non-trivial problem.

Example: Predicting working memory failure. Carlin *et al.* (1992) considered alternative predictions of human 'working memory' failure in computer-based tasks. (Working memory is a concept in cognitive psychology that has refined what was once called short-term memory.) Two alternative characterizations of working memory overload in database management query tasks involve (1) the number of conditions in the query and (2) the complexity of the query. To determine which characterization led to better predictions of error rates, two alternative statistical models were constructed based on alternative predictor variables, either the number of conditions or a measure of query complexity.

The database management system used in the study was SQL. An example of an SQL query is the following, in which information from a customer list is to be matched with that in an invoice list according to a criterion within the invoice list:

SELECT *name*
FROM *customer, invoice*
WHERE *amount* > 200
AND *invoice.id* = *customer.id*

This query would obtain the names of all customers who have any single invoice over 200 dollars. The last condition, called the 'join condition', is necessary to link the customer and invoice lists. If it were omitted, the results would be meaningless and the person attempting to generate the requested information would have to repeat the task. The focus of the experiment analyzed was omission of the join condition, which became a binary variable (omitted or not omitted).

There were 20 experimental subjects, each of whom were given 50 tasks to complete. Half the subjects received a 'cue', which reminded them to include the join condition, and half did not. (Interestingly, presence or absence of the cue turned out to have little effect on the probability of omitting the join condition.) Carlin *et al.* (1992) used a hierarchical logistic regression model which took the form

$$\text{logit}(p) = \text{subject effect within cue} + \text{effects of explanatory variables}$$

Here, the subject effect was modeled with a single parameter varying according to a normal distribution. The competing hypotheses involved competing explanatory variables: under the first, a single variable representing the number of conditions was used, while under the second, two variables characterized the complexity of the query. The two competing models had four parameters in common, which were considered *a priori* independent, leading to eight prior hyper-parameters (a location and scale for each parameter, with normal priors being used). In addition, the first model required two more hyper-parameters for the coefficient of the explanatory variable 'number of conditions' and the second required five hyper-parameters for the coefficients of the 'query complexity' variables (which were not assumed independent, so that a correlation hyper-parameter had to be introduced).

In total, then, there were 15 hyper-parameters that needed to be determined. The numerical problem involved 20 one-dimensional integrals nested within a five dimensional integral (for the first model) or a six-dimensional integral (for the second).

The Bayes factor was initially approximated using the Schwarz criterion, which indicated substantial evidence in favor of the query complexity model. Then, the 15 hyper-parameters for the priors were determined. Some of the information came from another related experiment, but some was based on rough guesswork. The Bayes factor was then computed using Laplace's method and was checked with subregion-adaptive integration. Sensitivity analysis was carried out, using the method described above, by shifting all prior means by one prior standard deviation unit in each direction, and then

doubling and halving each prior standard deviation. (Two seemingly appropriate alternative values were chosen for the one prior correlation, as well.) Thus, a total of about $3^{15} \approx 10^7$ alternative prior distributions were entertained. It was found that the common parameters and the explanatory-variable parameters were in fact observed-orthogonal in the sense of Section 5.2, and B was not very sensitive to the choice of prior on the common parameters. Some of the choices for the joint explanatory variable priors were eliminated as too extreme, and then over the remaining seemingly reasonable range the Bayes factors were found to be fairly consistent with the Schwarz criterion. The overall conclusion, taking into account the variation in the Bayes factors (and also the finding that a small number of subjects in the experiment carried a substantial portion of the comparative information), was that there was 'some evidence, though not strong evidence' in favor of the query complexity model.

8 Discussion

As I have tried to indicate, various advances in computational methods for Bayesian inference have made it possible to obtain Bayes factors in many practical problems. It is important, however, to keep in mind the sensitivity of results to the choice of priors that appear within the Bayes factor. In the face of this sensitivity, what is one to do?

The easiest way to proceed in practice is to ignore the problem by computing the Schwarz criterion and using it as a rough indication of the amount of evidence in favor of one or another hypothetical model. With sufficiently large samples this should be safe, by virtue of (5), in the sense that the same conclusion (based on orders of magnitude, as reviewed in Section 3) would be reached with the Schwarz criterion as with a full-blown subjective treatment of the kind described in Section 7. Furthermore, the result of Kass and Wasserman (1992b) indicates that for nested models the Schwarz criterion could also be considered a suitable 'reference' procedure, that is, it approximates the log of the Bayes factor obtained from a formal rule for selecting the prior. The difficulty is that, as in the example of Section 7, one may not know whether the sample size is indeed adequate for reliance on the Schwarz criterion, or for reliance on a formal rule for selecting priors. In this case, there seems to be no alternative to proceeding with subjective determination of the priors, which entails a substantial amount of further joint work between the statistician and the scientific collaborator.

It is perhaps of some consolation that computation for sensitivity analysis is relatively straightforward, but this still requires choices to be made for the priors. Thus, an important outstanding problem is to define a diagnostic that would allow a data analyst to determine, with minimal investment of additional effort, whether the Schwarz criterion (or, possibly, a Bayes factor with priors chosen by some formal rule) may be relied upon. In addition, while the advances in computation are encouraging, further progress is needed to bring the computational methods into the hands of those who are expert in using particular software implementations. Furthermore, though apparently useful suggestions have been made for computing Bayes factors using simulated samples from posterior distributions, a straightforward method that takes full advantage of the power and generality of Markov chain Monte Carlo simulation has not yet been devised.

References

BERGER, J. O. & DELAMPADY, M. (1987) Testing precise hypotheses, *Statistical Science*, 3, pp. 317–352.

CARLIN, B., KASS, R. E., LERCH, J. & HUGUENARD, B. (1992) Predicting working memory failure: a subjective Bayesian approach to model selection, *Journal of the American Statistical Association*, 87, pp. 319–327.

DELLAPORTAS, P. & WRIGHT, D. (1991) Positive embedded integration in Bayesian analysis, *Statistics and Computing*, 1, pp. 1–12.

GELFAND, A. & DEY, D. (1993) *Bayesian Model Choice: Asymptotics and Exact Calculations*, Technical Report, Department of Statistics, University of Connecticut.

GENZ, A. & KASS, R. E. (1993) *Subregion Adaptive Integration of Functions Having a Dominant Peak*, Technical Report, Department of Statistics, Carnegie-Mellon University.

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

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

KASS, R. E. & GREENHOUSE, J. B. (1989) Comment on 'investigating therapies of potentially great benefit: ECMO', by Ware (1989), *Statistical Science*, 4, pp. 310–317.

KASS, R. E. & RAFTERY, A. E. (1993). *Bayes Factors*, Technical Report, Department of Statistics, Carnegie-Mellon University.

KASS, R. E. & VAIDYANATHAN, S. (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions, *Journal of the Royal Statistical Society, Series B*, 54, pp. 129–144.

KASS, R. E. & WASSERMAN, L. (1992a) *Improving the Laplace Approximation Using Posterior Simulation*, Technical Report, Department of Statistics, Carnegie-Mellon University.

KASS, R. E. & WASSERMAN, L. (1992b) *A Reference Bayesian Test of Nested Hypotheses with Large Samples*, Technical Report, Department of Statistics, Carnegie-Mellon University.

LINDLEY, D. V. (1957) A statistical paradox, *Biometrika*, 44, pp. 187–192.

NAYLOR, J. C. & SMITH, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions, *Applied Statistics*, 31, pp. 214–225.

NEWTON, M. A. & RAFTERY, A. E. (1991) *Approximate Bayesian Influence by the Weighted Likelihood Bootstrap*, Technical Report, Department of Statistics, University of Washington.

SCHWARZ, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, pp. 461–464.

SKLAR, R. & STRAUSS, B. (1980) Role of the *uvrE* gene product and of inducible 0*-methylguanine removal in the induction of mutations by *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine in *Escherichia coli, Journal of Molecular Biology*, 143, pp. 345–363.

TIERNEY, L. & KADANE, J. B. (1986) Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, 81, pp. 82–86.

TIERNEY, L., KASS, R. E. & KADANE, J. B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association*, 84, pp. 710–716.

WARE, J. H. (1989) Investigating therapies of potentially great benefit: ECMO (with discussion), *Statistical Science*, 4, pp. 298–340.