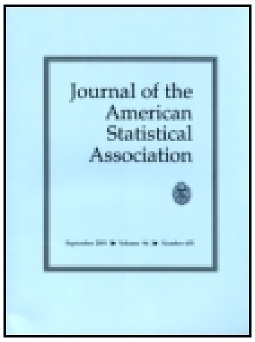

Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion

Robert E. Kass & Larry Wasserman

A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion

Robert E. KASS and Larry WASSERMAN*

To compute a Bayes factor for testing H_0: $\psi = \psi_0$ in the presence of a nuisance parameter β, priors under the null and alternative hypotheses must be chosen. As in Bayesian estimation, an important problem has been to define automatic, or "reference," methods for determining priors based only on the structure of the model. In this article we apply the heuristic device of taking the amount of information in the prior on ψ equal to the amount of information in a single observation. Then, after transforming β to be "null orthogonal" to ψ, we take the marginal priors on β to be equal under the null and alternative hypotheses. Doing so, and taking the prior on ψ to be Normal, we find that the log of the Bayes factor may be approximated by the Schwarz criterion with an error of order $O_p(n^{-1/2})$, rather than the usual error of order $O_p(1)$. This result suggests the Schwarz criterion should provide sensible approximate solutions to Bayesian testing problems, at least when the hypotheses are nested. When instead the prior on ψ is elliptically Cauchy, a constant correction term must be added to the Schwarz criterion; the result then becomes a multidimensional generalization of Jeffreys's method.

KEY WORDS: Bayes information criterion; Laplace's method; Model selection; Null-orthogonal parameters; Orthogonal parameters.

1. INTRODUCTION

Bayesian tests are carried out using Bayes factors, which require prior distributions on the parameters appearing in the null and alternative models. In principle, priors formally represent available information, but in practice automatic or "reference" procedures are often used; these are computed from modeling assumptions and do not depend otherwise on specifics of the problem. For estimation problems, reference priors are often "flat" (uniform) in some sense on parameters of interest but in testing such a prescription leads to serious difficulties. (See Kass and Wasserman 1995c for a survey of the literature on reference priors and Kass and Raftery 1995 for a review of Bayes factors.) Thus an important problem is to define a reference Bayesian testing procedure that uses a proper prior on the parameter of interest. This article presents such a procedure, which we find intuitively reasonable, and then shows that it leads to the Schwarz criterion (Schwarz 1978) and also, with a simple modification, generalizes a proposal of Jeffreys (1961, chap. 5). We also connect the idea with work of Smith and Spiegelhalter (1980) and Zellner and Siow (1980).

Let $Y = (Y_1 \ldots, Y_n)$ be iid observations from a family parameterized by (β, ψ), with $\dim(\beta, \psi) = m$ and $\dim(\beta) = m_0$. The hypothesis H_0: $\psi = \psi_0$ is to be tested against the alternative H_1: $\psi \in \mathbb{R}^{m-m_0}$ using a Bayes factor,

$$B = \frac{\int p(y \mid \beta, \psi_0)\pi_0(\beta)\, d\beta}{\int p(y \mid \beta, \psi)\pi(\beta, \psi)\, d\beta\, d\psi},$$

where $p(y \mid \beta, \psi)$ denotes the probability density for the data and $\pi_0(\beta)$ and $\pi(\beta, \psi)$ are the priors under the null and alternative hypotheses. The Schwarz criterion is

$$S = l_0(\hat{\beta}_0) - l(\hat{\beta}, \hat{\psi}) + \frac{1}{2}(m - m_0)\log n,$$

where $\hat{\beta}_0$ maximizes of the null-hypothetical log-likelihood $l_0(\beta) = \log p(y \mid \beta, \psi_0)$ and $(\hat{\beta}, \hat{\psi})$ maximizes the unrestricted log-likelihood $l(\beta, \psi)$.

The Schwarz criterion (also known as the Bayes information criterion or BIC, though in the usual definition $2S$ = BIC) is well established in the literature on model selection (Smith and Spiegelhalter 1980)—especially in time series analysis (Hannan 1980), where its asymptotic consistency separates it from some of its competitors. It achieves this consistency by crudely approximating the log Bayes factor, which is necessarily asymptotically consistent under fairly general conditions. For example, from Doob's theorem (Doob 1949), the consistency of the Bayes factor follows easily, almost surely with respect to the prior. The approximation ignores terms of constant order, including those arising from the prior. This gives

$$\frac{\log(B) - Sp}{\log(B)} \rightarrow 0,$$

which suffices for consistency, but the crudeness of the approximation allows

$$\frac{\exp(S)}{B} \not\rightarrow 1;$$

that is, at least for some priors, $\exp(S)$ will be a poor approximation to the Bayes factor and thus a dubious quantification of evidence in favor of a model. Our main result is that an intuitive reasonable choice of priors leads to $\exp(S)/B \rightarrow 1$, with error of order $O_p(n^{-1/2})$; see also Stone (1979). This says that for large samples, the Schwarz criterion, when exponentiated, provides an interesting approximate Bayes factor and thus a potentially useful quantification of evidence. On the other hand, for model selection, the result gives a calibration for the Schwarz criterion (in terms of

* Robert E. Kass and Larry Wasserman are Professors, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. The authors thank Jim Berger, Donna Pauler, Adrian Raftery, and the referees for their comments. Both authors were supported by National Science Foundation Grant DMS-9303557 and National Institutes of Health Grant R01-CA54852-01; Wasserman was also supported by National Science Foundation Grant DMS-9357646.

Journal of the American Statistical Association
September 1995, Vol. 90, No. 431, Theory and Methods

posterior probability), so that "large" values may be distinguished from unremarkable values.

An especially simple special case of the method illustrates the basic idea. Let $Y_i \sim N(\psi, \sigma^2)$, iid for $i = 1, \ldots, n$, with σ known, the hypotheses to be tested being H_0: $\psi = \psi_0$ versus H_1: $\psi \in \mathbb{R}$. The prior distribution on ψ under H_1,

$$\psi \sim N(\psi_0, \tau^2) \quad \text{with} \quad \tau = \sigma,$$

has the interpretation that "the amount of information in the prior on ψ is equal to the amount of information about ψ contained in one observation." This is precisely what we find intuitively appealing. Furthermore, an easy calculation shows that the resulting logarithm of the Bayes factor is approximated by the Schwarz criterion with error of order $O_p(n^{-1/2})$. In Section 2 we generalize to arbitrary regular parametric families using Fisher information to define "amount of information" and suitably transforming any nuisance parameters.

It would be possible (and may be desirable) to use non-Normal priors with the scheme that we propose. We note in particular that if a Cauchy prior is used, then a correction must be added to the Schwarz criterion; this results in a generalization of Jeffreys's proposal (Jeffreys 1961), which we describe in Section 3.

The conclusion that we draw from the results in Sections 2 and 3, together with the examples in Section 4, is that there is good motivation for using the Schwarz criterion, or some modification of it, as a large-sample testing procedure. In addition, the main result explains what has sometimes seemed a surprisingly good agreement between the Schwarz criterion and Bayes factors computed with subjectively determined priors (e.g., Carlin, Kass, Lerch, and Huguenard 1992; Raftery 1993). In these cases the subjectively determined priors are not very different from those that take the information in the prior to be equal to that in one observation.

2. RESULT USING NORMAL PRIORS

The results presented here concern iid data as discussed in Section 1 (though see Sec. 5 for remarks on generality) and require two important simplifying assumptions and some regularity conditions. First, we assume that β and ψ are null orthogonal, meaning that the Fisher information matrix $I(\beta, \psi)$ is block diagonal for null hypothetical parameter values; that is, $I_{\beta\psi}(\beta, \psi_0) = 0$ for all β. As noted by Kass and Vaidyanathan (1992), it is always possible to transform β so that it becomes null orthogonal to ψ. Second, we assume that the marginal prior on β is the same under both hypotheses; that is,

$$\pi_0(\beta) = \int \pi(\beta, \psi)\, d\psi.$$

We write the marginal prior on ψ under H_1 as $\pi_\psi(\psi)$. For now we assume, in addition, that β and ψ are a priori independent under the alternative but will comment on this at the end of the section.

As far as regularity conditions are concerned, we assume that Laplace's method may be applied in both the numerator and denominator of the Bayes factor (the models are "Laplace regular" in the terminology of Kass, Tierney, and Kadane 1990) and also

$$-n^{-1}D^2 l(\hat{\beta}, \hat{\psi}) - I(\beta, \psi) = O_p(n^{-1/2}), \tag{1}$$

where $(\hat{\beta}, \hat{\psi})$ is the maximum likelihood estimator (MLE). Finally, we perform the computation with the assumption that the MLE $\hat{\psi}$ under the alternative satisfies $\psi_0 - \hat{\psi} = O_p(n^{-1/2})$ as it would if the "true" value of ψ were either ψ_0 or a neighboring alternative ψ_n, such that $\psi_n - \psi_0 = O(n^{-1/2})$. As Kass and Vaidyanathan (1992) noted, when this situation does not hold, the Bayes factor is exponentially small, almost surely, and will quickly become decisive for large samples, and it will no longer make much practical difference whether an approximation is accurate or crude.

Under the foregoing conditions, Kass and Vaidyanathan obtained the approximation to the Bayes factor,

$$B = (2\pi)^{(m_0-m)/2} \frac{|\Sigma_0|^{1/2}}{|\Sigma|^{1/2}} \exp\{l(\hat{\beta}_0, \psi_0) - l(\hat{\beta}, \hat{\psi})\} \times \frac{1}{\pi_\psi(\hat{\psi})} \left\{1 + O_p\!\left(\frac{1}{n}\right)\right\}, \tag{2}$$

where $\Sigma_0 = (-D^2 l_0(\hat{\beta}_0))^{-1}$ and $\Sigma = (-D^2 l(\hat{\beta}, \hat{\psi}))^{-1}$, with $\hat{\beta}_0$ defined as in Section 1 to be the MLE under H_0. (Equation (2) followed from the observation that the likelihood equations are approximately separable in the two variables, so that the MLE's of β under the two hypotheses are close; specifically, $\hat{\beta}_0 - \hat{\beta} = O_p(n^{-1})$.) Notice that the approximation in (2) does not involve the prior on β. This was the main motivation of Kass and Vaidyanathan for transforming β so that it becomes null orthogonal to ψ; see also Cox & Hinkley (1980 pp. 160–162).

The following proposition drives the main result.

Proposition. Under the regularity conditions outlined previously, as $n \to \infty$,

$$B = e^S \cdot \{(2\pi)^{(m_0-m)/2} |I_{\psi\psi}(\beta, \psi_0)|^{-1/2} \pi_\psi(\hat{\psi})\}^{-1} \times \{1 + O_p(n^{-1/2})\}. \tag{3}$$

Proof. From (1) and the block diagonality of $I(\beta, \psi_0)$ it follows that

$$n^{(m_0-m)/2} \det(-D^2 l_0(\hat{\beta}))^{-1/2} \det(-D^2 l(\hat{\beta}, \hat{\psi}))^{1/2} = \det(I_{\psi\psi}(\beta, \psi_0))^{1/2}(1 + O_p(n^{-1/2})).$$

Putting this in Equation (2) yields (3).

We now specialize to the case in which the prior on ψ under H_1 is elliptically symmetric having location ψ_0 and scale matrix Σ_ψ, with density $\pi_\psi(\psi) = |\Sigma_\psi|^{-1/2} f((\psi - \psi_0)^T \Sigma_\psi^{-1} (\psi - \psi_0))$ for some $m - m_0$-dimensional multivariate density f. When f is the standard multivariate Normal density, we have $\psi \sim N(\psi_0, \Sigma_\psi)$. We choose Σ_ψ by generalizing the requirement given in Section 1 that the amount of information in the prior be equal to the amount of information in one observation. We take Σ_ψ to satisfy

$$|\Sigma_\psi|^{-1} = |I_{\psi\psi}(\beta, \psi_0)|, \tag{4}$$

where $I_{\psi\psi}(\beta, \psi_0)$ is the block of $I(\beta, \psi_0)$ corresponding to ψ. (By null orthogonality, under (β, ψ_0), the matrix $I_{\psi\psi}(\beta, \psi_0)$ is the inverse of the asymptotic variance matrix of the MLE $\hat{\psi}$.)

Result. Under the conditions leading to the foregoing proposition, if the prior on ψ is elliptically symmetric with density $\pi_\psi(\psi) = |\Sigma_\psi|^{-1/2} f((\psi - \psi_0)^T \Sigma_\psi^{-1}(\psi - \psi_0))$, where Σ_ψ satisfies (4), then as $n \to \infty$,

$$B = e^S \cdot \{(2\pi)^{(m_0 - m)/2} f(\hat{\psi})\}^{-1}\{1 + O_p(n^{-1/2})\}. \quad (5)$$

In the special case in which the prior on ψ is $N(\psi_0, \Sigma_\psi)$,

$$\log B = S + O_p(n^{-1/2}). \quad (6)$$

Proof. Equation (5) is an immediate consequence of the proposition, and (6) then follows from $\pi_\psi(\hat{\psi}) = \pi_\psi(\psi_0)\{1 + O_p(n^{-1/2})\}$.

These results show how null-orthogonality and (4) greatly simplify the Bayes factor; when the prior is Normal, its log is approximately equal to the Schwarz criterion; otherwise a correction term must be added to obtain a similarly accurate approximation. In the next section we consider the latter situation for an elliptically Cauchy distribution. We call priors based on (4) *unit information* priors.

Before leaving this section we note that in these main results, (5) and (6), we have assumed ψ to be Normal according to (4) independently of β. But it can happen that $I_{\psi\psi}(\beta, \psi)$ depends on β. In this case we must assume that $\pi_\psi(\psi)$ specifies the *conditional* distribution of ψ given β rather than the marginal distribution. The results then continue to hold.

3. JEFFREYS'S METHOD

For Normal location testing problems, Jeffreys (1961, pp. 268–270) argued that the prior should be symmetric and should have no moments. He considered the Cauchy density to be the "simplest function" satisfying his requirements and thus chose it for the prior. He then applied his argument to general additional steps, which we outline. We then note that when the prior in (5) is elliptically Cauchy, the result (5) furnishes a multivariate generalization of Jeffreys's method.

In treating Bayes factors generally, Jeffreys (1961, p. 249) assumed that they were globally orthogonal, meaning that $I(\beta, \psi)$ was diagonal for all (β, ψ). (Kass and Vaidyanathan (1992) used null-orthogonality because it is not always possible to produce globally orthogonal parameters in multidimensional cases.) Jeffreys (pp. 275 and 277) took the prior under the alternative to be Cauchy in terms of the square root of the symmetrized Kullback–Leibler number, which we may write here as $J = K((\beta, \psi), (\beta, \psi_0)) + K((\beta, \psi_0), (\beta, \psi))$, so that

$$\pi_\psi(\psi) d\psi = \frac{1}{\pi} \frac{|dJ^{1/2}|}{1 + J}, \quad (7)$$

J being a function of ψ for each given β. He also (essentially, p. 277) used the approximation

$$K((\beta, \psi), (\beta, \psi_0)) + K((\beta, \psi_0), (\beta, \psi)) = I_{\psi\psi}(\beta, \psi_0)(\psi - \psi_0)^2 + O(|\psi - \psi_0|^3) \quad (8)$$

for computational purposes.

The intuition behind (7) comes from Jeffreys's having recognized that his method for Normal location problems could be interpreted as putting a Cauchy prior on what we might call the "distance from the null model," with "distance" measured by the square root of the Kullback–Leibler number. (He had already established that the symmetrized Kullback–Leibler number behaves like a squared distance function locally (see Kass 1989 for additional geometrical discussion).

For multidimensional ψ, the apparent generalization is to use an elliptically Cauchy prior that is uniform on all points equally "distant" from the model with its scale factor determined by the Kullback–Leibler "distance." Substituting the multidimensional version of (8), in which the right-hand side becomes $(\psi - \psi_0)^T I_{\psi\psi}(\beta, \psi_0)(\psi - \psi_0) + O(\|\psi - \psi_0\|^3)$, this prescription amounts to taking an elliptically Cauchy prior on ψ centered at ψ_0 with scale matrix $I_{\psi\psi}(\beta, \psi_0)^{-1}$. Thus, by replacing Jeffreys's requirement of orthogonal parameters with the weaker (and always possible) specification that β be null-orthogonal to ψ, we obtain the following as a multidimensional generalization of Jeffreys's approximate method for testing hypotheses.

Result. In the special case of (5) in which the prior is elliptically Cauchy, centered at ψ_0 with scale matrix Σ_ψ satisfying (4),

$$\log B = S_C + O_p(n^{-1/2}), \quad (9)$$

where $S_C = S - \log r$ with r being the ratio of the $(m - m_0)$-dimensional spherical Cauchy and Normal densities at the origin; that is,

$$r = 2^{(m - m_0)/2} \Gamma\left(\frac{m - m_0 + 1}{2}\right) \Big/ \sqrt{\pi}.$$

Proof. Immediate from (5) and $\pi_\psi(\hat{\psi}) = \pi_\psi(\psi_0)[1 + O(n^{-1/2})]$.

4. EXAMPLES

Here we consider several examples to illustrate the accuracy of the approximations (6) and (9). The first two compare S and S_C to the exact Bayes factors using Normal and Cauchy priors in one-dimensional Normal and Cauchy models. The third example compares S to the exact Bayes factor in a multivariate Normal. In the fourth example we compare the Schwarz approximation and the exact Bayes factor when performing model selection in a probit regression setting. The assumptions of the results in Section 2 are violated in the fourth example; nonetheless, the approximation turns out to be quite accurate.

Example 1: Normal Model. We begin by returning to the simple motivating example discussed in Section 1. Let $Y \sim N(\psi, \sigma^2/n)$ with σ known. Without loss of generality, we take $\sigma = 1$ and consider the test of H_0: $\psi = 0$ versus H_1: $\psi \in \mathbb{R}$. The Normal unit-information prior, used in (6), under H_1 is then $\psi \sim N(0, 1)$. The exact log Bayes factor is

$$\log B = \frac{1}{2}\log(n + 1) - \frac{ny^2}{2}\frac{n}{n + 1},$$

whereas the Schwarz approximation is

$$S = \frac{1}{2}\log(n) - \frac{ny^2}{2}.$$

The plots in the first column of Figure 1 show these two quantities when $n = 5$ and $n = 25$ for several values of y. The plots in the second column of Figure 1 show the corresponding quantities when S_C is used. The plots are on a $\log_{10}$ scale. The approximation is very accurate even for small sample sizes and even when the observed value is more than three standard errors form the null (i.e., even when $|y| > 3/\sqrt{n}$).

Example 2: Cauchy Model. Now suppose that $Y_1, \ldots, Y_n$ are iid observations with a Cauchy(ψ, 1) distribution, the hypotheses being H_0: $\psi = 0$ versus H_1: $\psi \in \mathbb{R}$. The Normal unit-information reference prior under H_1 using (4) is then $\psi \sim N(0, 2)$. Figure 2 shows the results from some simulated data sets plotted as exact Bayes factor B versus the Schwarz criterion S. The solid line is $B = S$. The top two rows are based on the $N(0, 2)$ prior; the bottom two rows correspond to the Cauchy(0, 2) prior, which is the prior used in (9), together with S_C in place of S. The columns correspond to $\psi = 0, 1, 2$.

The results show that for n as small as 5, the approximations are reasonably good. For $n = 25$, the agreement between exact and approximate values is excellent even when ψ is far from the null hypothesis.

Example 3: Multivariate Normal. Suppose now that $Y_1, \ldots, Y_k \sim N_p(\psi, I)$, where I is a p by p identity matrix and

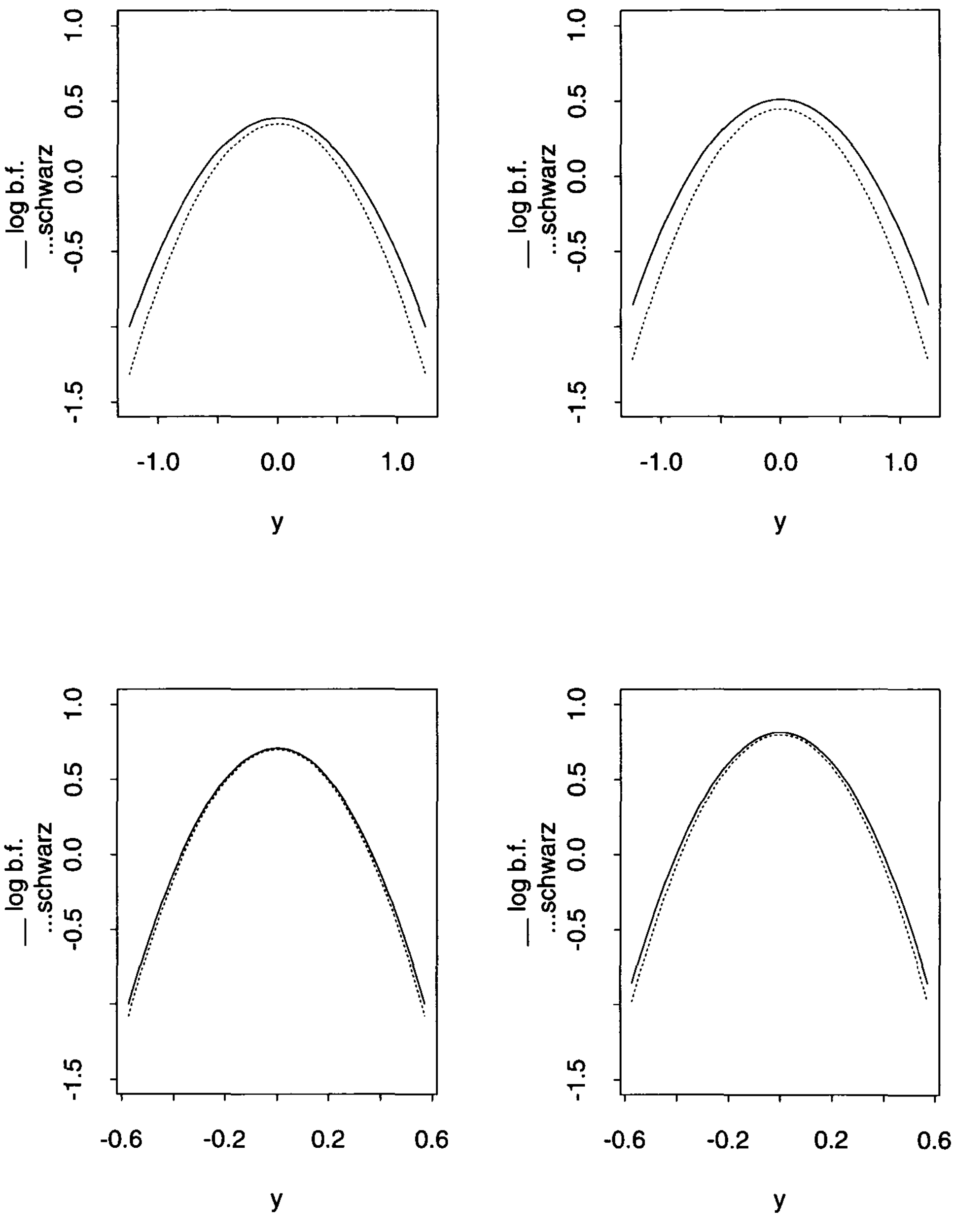

Figure 1. The Log Bayes Factor (Solid Line) and Schwarz Approximation (Dotted Line) as a Function of y ~ N(θ, 1/n). The two plots on the left correspond to a normal prior; the two plots on the right correspond to a Cauchy prior. The sample sizes are n = 5 for the two plots on the top row and n = 25 for the two plots on the bottom row.

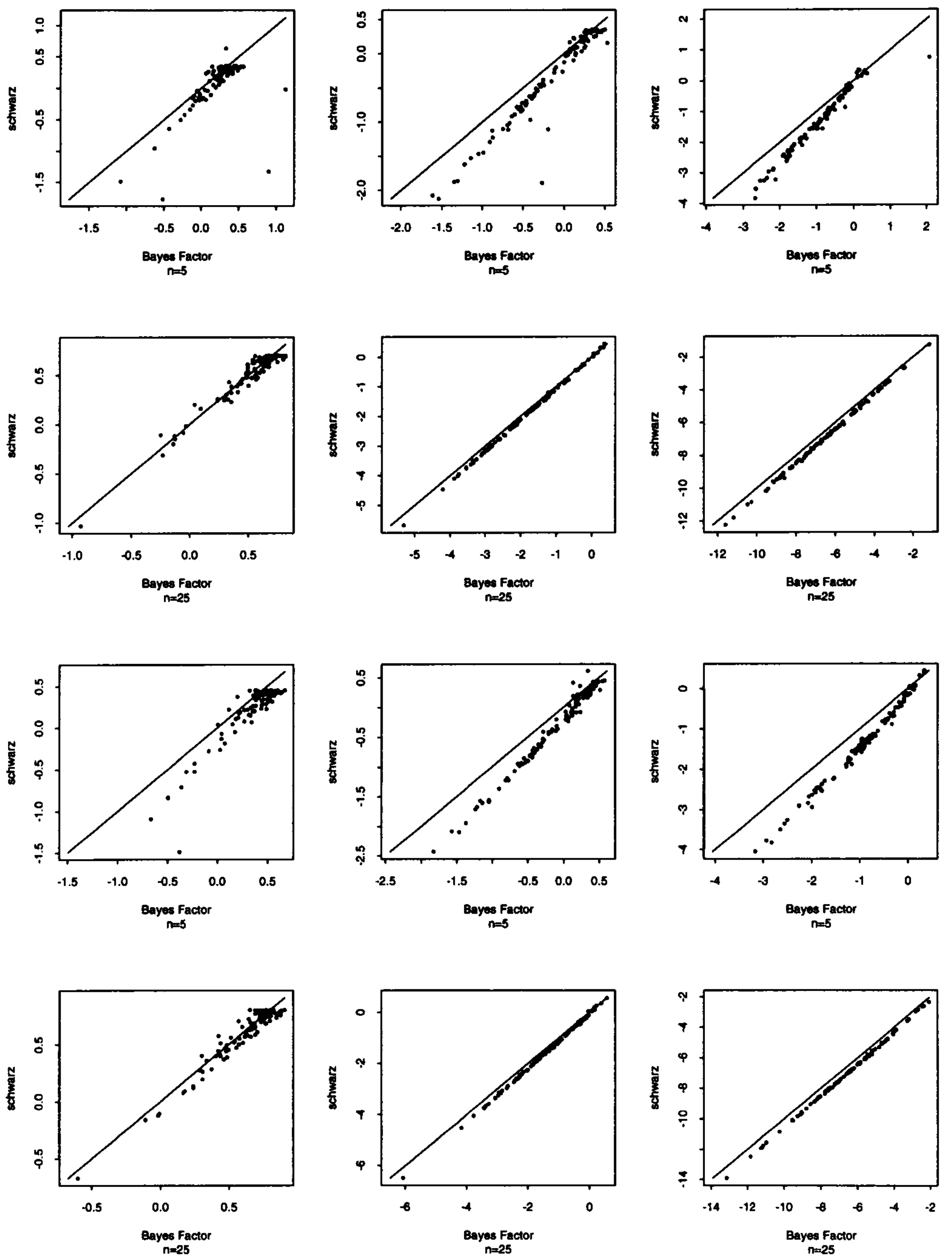

Figure 2. Plots of Schwarz Approximation Versus Log Bayes Factor From Several Simulations. The data are from a Cauchy distribution with median θ. The three columns of plots correspond to $\theta = 0, 1, 2$. The first two rows of plots are based on normal priors; the second two rows are based on Cauchy priors. Points near the diagonal represent cases where the Schwarz approximation is accurate.

we wish to test H_0: $\psi = 0$ versus H_1: $\psi \in \mathbb{R}^p$. If we identify one unit of information as a single scalar observation so that each vector Y_i contains p units of information, then the reference prior is $N_p(0, pI)$, the Bayes factor is

$$B = (n+1)^{p/2}\exp\left\{-\frac{k}{2}\,|\bar{y}|^2\frac{n}{n+1}\right\}$$

where $n = kp$, and the Schwarz criterion is

$$\exp\{S\} = n^{p/2}\exp\left\{-\frac{k}{2}\,|\bar{y}|^2\right\}.$$

The ratio R is given by

$$R = \left(1+\frac{1}{n}\right)^{p/2}\exp\left\{\frac{k|y|^2}{2(n+1)}\right\}$$
$$= n^{p/(2n)}\left(1+\frac{1}{n}\right)^{p(n+1)/(2n)}B^{-1/n}.$$

Table 1 gives R for various k, p, and B. Again we see that S is quite accurate. (Note that we need $B \leq (n+1)^{p/2}$; otherwise, there is no point y in the sample space that gives that value of B. But for completeness, we have computed R for all combinations of k, p, and B in the table.)

Table 1. The Ratio $R = B/\exp(S)$ as a Function of Dimension (p), Replications (k), and B

	p = 1					p = 10				
	B = 1/10	B = 1/3	B = 1	B = 3	B = 10	B = 1/10	B = 1/3	B = 1	B = 13	B = 10
5	2.1	1.6	1.3	1.1	0.8	1.7	1.7	1.6	1.6	1.6
10	1.5	1.3	1.2	1.1	0.9	1.4	1.3	1.3	1.3	1.3
25	1.2	1.1	1.1	1.0	1.0	1.1	1.1	1.1	1.1	1.1
100	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Example 4: Probit Regression. Suppose that $Y_1, \ldots, Y_n$ are binary random variables and that $p_i = \Pr(Y_i = 1 | x_i) = \Phi(x_i' \beta)$, where Φ is the standard Normal cdf, x_i is a $K \times 1$ vector of covariates, and β is a $K \times 1$ vector of unknown parameters. We consider a data set from Finney (1947), where Y_i corresponds to presence ($Y_i = 1$) or absence ($Y_i = 0$) of vasoconstriction; Bayesian model selection for these data was discussed by Raftery (1993). The covariate $x_1 \equiv 1$ is an intercept, x_2 is rate of inspiration, and x_3 is volume of air inspired. First, consider testing H_1: ($\beta_2 = 0, \beta_3 = 0$) versus H_{123}: $(\beta_2, \beta_3) \in \mathbb{R}^2$, so that our ψ becomes $\psi = (\beta_2, \beta_3)$. The subscript on H indicates which variables are included in the regression. We note that, technically this example violates the assumptions of the result (6), because the data are not identically distributed and the information is not null-orthogonal. We reduce the effect of the latter by centering all the covariates so that the expected information matrix based on all the data is

$$nI = \begin{bmatrix} .072 & .044 & .010 \\ .04 & .51 & .23 \\ .010 & .23 & .21 \end{bmatrix}.$$

We write this as nI, because the per unit information is taken to be $1/n$ times this matrix. We take β_0 to have a flat prior and ψ to have the reference Normal distribution with covariance matrix $\Sigma_\psi = I_{\psi\psi}^{-1}(\beta, \psi_0)$ where $I_{\psi\psi}$ represents the 2 by 2 information matrix for ψ and $n = 37$ is the number of observed Y_i''s.

It is difficult to compute an exact answer here, so we resorted to the following Monte Carlo method. First, we drew a sample from the posterior using Gibbs sampling as described by Albert and Chib (1993). Following Kass and Wasserman (1992) and Raftery (1994), we used the sample to estimate the Laplace approximation to the Bayes factor. We also considered several corrections proposed by Kass and Wasserman and by DiCiccio, Kass, Raftery, and Wasserman (1995). All these methods gave approximately the same value as the simulated Laplace method. (The calculations were based on 10,000 draws from the posterior; the answer did not change appreciably after 1000 draws.) This value, along with the Schwarz approximation and some other model comparisons are given in Table 2. In each, the approximation is very accurate.

5. DISCUSSION

The main conclusion we draw from our results and numerical comparisons is that the Schwarz criterion S (or some simple modification such as S_C) furnishes an interesting approximately Bayesian testing procedure. The point is that S is easy to compute and does not require explicit introduction of prior distributions into its calculation, and moreover, the implicit priors that make it approximately a log Bayes factor [according to (6)] are intuitively reasonable and the sample sizes needed to provide accuracy of the approximation are not prohibitively large. We thus find the Schwarz criterion a useful "automatic" Bayesian testing procedure for nested models. As such, we believe that it may be preferable to the intrinsic Bayes factors of Berger and Pericchi (1993, 1995) and the fractional Bayes factors of O'Hagan (1995), although in some cases these approaches will yield similar results (see Kass and Wasserman (1995a, 1995b).

Several issues deserve further comment. First, an important practical point concerning use of S is that the sample size n appearing in the formula must be determined carefully. It is apparent from the derivation of (6) using (1) that n should be the rate at which the Hessian matrix of the log-likelihood function grows; thus n becomes the number of data values contributing to the summation that appears in the formula for the Hessian. In our multivariate Normal example, for instance, we took $n = kp$, so that n is the total number of *scalar* observations. In doing so we have actually substituted $(1/p)I(\psi)$ for $I(\psi)$ in (1). An alternative would be to take $n = k$ (which, for fixed p, is of the same order as kp). We hope to elaborate on this subtle but important matter elsewhere.

Second, the analytical results presented here apply fairly generally to iid random variables, requiring standard regularity conditions for asymptotic expansions. These include the restriction that the MLE lies in the interior of the parameter space. Occasionally problems arise in which the MLE is on the boundary of the parameter space (e.g., when a variance component has an MLE of zero). Hsiao (1994) discussed this case and used a

Table 2. Comparison of $\log_{10} B$ and S_{10} for the Data From Finney (1947)

Models	$\text{Log}_{10} B$	S_{10}
M_1 vs. M_{123}	−3.60	−3.63
M_1 vs. M_{12}	−.76	−.77
M_1 vs. M_{13}	−.17	−.17

NOTE: Where $S_{10} = \log_{10} e \cdot S$ is the Schwarz criterion expressed in terms of log and base-10.

modified version of the method presented here in treating the problem of testing for extrabinomial variability; the new approximation is again quite accurate in the examples that Hsiao treated.

Strictly speaking, our results do not actually apply to special cases, such as in linear and generalized linear models, when the sampling is not iid. The heuristics appear sound, however, and we are confident that a rigorous extension is possible in such situations; we hope to have some further results on this in the future (following the approach of Kass et al. 1990). In fact, the accuracy of the approximation (6) in the probit regression example indicates its applicability in non-iid settings—indeed, in settings in which exact null-orthogonality also does not hold.

In the special case of Normal linear models with known variance and orthogonal design matrices, Smith and Spiegelhalter (1980) gave a version of result (6) together with the appealing interpretation of the prior that we have emphasized here. (Orthogonal design matrices entail orthogonality of the parameters, which satisfies the assumption of null-orthogonality used here; it should be noted that the design matrices can always be made orthogonal by transformation without changing the nested testing problem.) A closely related alternative to our S_C, again in the special case of Normal linear models with orthogonal design matrices but allowing the variance to be unknown, was given by Zellner and Siow (1980). Their approximation was also intended to furnish a generalization of Jeffreys's method. It has the same order of accuracy as (9) but involves an analytical integration over the unknown variance. This brings up the additional important point that there are simple alternatives to S and S_C that might be of interest. For instance, one could apply an approximation after integrating out some parameter, as in the work of Zellner and Siow (1980), who followed Jeffreys (1961), or one could use a different estimator in place of the MLE (such as the usual "unbiased" estimator of variance); one could also replace the value of the prior density at ψ_0 used in (6) and (9) with its value at $\hat{\psi}$ as in (5). Whether there are important numerical advantages to such modifications is another topic for future research.

Finally, the Schwarz criterion is well defined for nonnested models. There are thus, in that case, the dual open questions of whether a reference Bayesian test similar to the one used here may be formulated, and how, if at all, the Schwarz criterion would have to be modified to become an approximate Bayes factor to order $O_p(n^{-1/2})$. We expect to report work on this problem in a future paper.

[Received September 1993. Revised August 1994.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association,* 88, 669–679.

Berger, J. O., and Pericchi, L. R. (1993), "The Intrinsic Bayes Factor for Model Selection and Prediction," Technical Report 93-43C, Purdue University, Dept. of Statistics.

——— (1995), "The Intrinsic Bayes Factor for Linear Models," in *The Proceedings of the Fifth Valencia Conference on Bayesian Statistics.*

Cox, D. R., and Hinkley, D. V. (1980), *Problems and Solutions in Theoretical Statistics.* London: Chapman and Hall.

DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L. (1995), "Computing Bayes Factors by Combining Simulation and Large Sample Approximations," manuscript in progress.

Doob, J. L. (1949), "Application of the Theory of Martingales," in *Colloques Internationaux du Centre National de la Recherche Scientifique,* Paris, pp. 23–27.

Finney, D. J. (1947), "The Estimation From Individual Records of the Relationship Between Dose and Quantal Response," *Biometrika,* 34, 320–334.

Hannan, E. J. (1980), "The Estimation of the Order of an ARMA Process," *The Annals of Statistics,* 8, 1071–1081.

Hsiao, C. K. (1994), "Bayesian Tests of Extra Binomial Variability With Emphasis on the Boundary Case," Ph.D. dissertation, Carnegie Mellon University, Dept. of Statistics.

Kass, R. E. (1989), "The Geometry of Asymptotic Inference" (with discussion), *Statistical Science,* 4, 188–234.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors and Model Uncertainty," *Journal of the American Statistical Association,* 90, 773–795.

Kass, R. E., Tierney, L., and Kadane, J. B. (1990), "The Validity of Posterior Expansions Based on Laplace's Method," in *Essays in Honor of George Barnard,* eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland, pp. 473–488.

Kass, R. E., and Vaidyanathan, S. (1992), "Approximate Bayes Factors and Orthogonal Parameters, With Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society,* Ser. B, 54, 129–144.

Kass, R., and Wasserman, L. (1992), "Improving the Laplace Approximation Using Posterior Simulation," Technical Report 566, Carnegie Mellon University, Dept. of Statistics.

——— (1995a), Comment on O'Hagan's "Fractional Bayes Factors for Model Comparisons," *Journal of the Royal Statistical Society,* Ser. B, 57, 131–132.

——— (1995b), Comment on Berger and Pericchi's "The Intrinsic Bayes Factor for Linear Models," in *Proceedings of the Fifth Valencia Conference on Bayesian Statistics.*

——— (1995c), "Formal Rules for Constructing Priors," *Journal of the American Statistical Association,* to appear.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparisons," *Journal of the Royal Statistical Society,* Ser. B, 57, 99–138.

Raftery, A. E. (1993), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," Technical Report 255, University of Washington, Dept. of Statistics.

Raftery, A. E. (1994), "Hypothesis Testing and Model Selection Via Posterior Simulation," to appear in *Practical Markov Chain Monte Carlo,* eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461–464.

Smith, A. M. F., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society,* Ser. B, 42, 213–220.

Stone, M. (1979), "Comments on Model Selection Criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society,* Ser. B, 41, 276–278.

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: University of Valencia Press.