Computing Bayes Factors by Combining Simulation and Asymptotic Approximations

Thomas J. DiCiccio; Robert E. Kass; Adrian Raftery; Larry Wasserman

Journal of the American Statistical Association, Vol. 92, No. 439 (Sep., 1997), 903-915.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199709%2992%3A439%3C903%3ACBFBCS%3E2.0.CO%3B2-G

Computing Bayes Factors By Combining Simulation and Asymptotic Approximations

Thomas J. DiCiccio, Robert E. Kass, Adrian Raftery, and Larry Wasserman

The Bayes factor is a ratio of two posterior normalizing constants, which may be difficult to compute. We compare several methods of estimating Bayes factors when it is possible to simulate observations from the posterior distributions, via Markov chain Monte Carlo or other techniques. The methods that we study are all easily applied without consideration of special features of the problem, provided that each posterior distribution is well behaved in the sense of having a single dominant mode. We consider a simulated version of Laplace's method, a simulated version of Bartlett correction, importance sampling, and a reciprocal importance sampling technique. We also introduce local volume corrections for each of these. In addition, we apply the bridge sampling method of Meng and Wong. We find that a simulated version of Laplace's method, with local volume correction, furnishes an accurate approximation that is especially useful when likelihood function evaluations are costly. A simple bridge sampling technique in conjunction with Laplace's method often achieves an order of magnitude improvement in accuracy.

KEY WORDS: Bartlett corrections; Laplace's method; Model selection; Monte Carlo.

1. INTRODUCTION

Recently developed methods for simulating observations from posterior distributions greatly enhance the applicability of Bayesian inference. These methods include Markov chain Monte Carlo methods (Besag, Green, Higdon, and Mengersen 1995; Smith and Roberts 1992; Tierney 1994), the sampling importance resampling (SIR) algorithm (Rubin 1987, 1988), and the weighted likelihood bootstrap (Newton and Raftery 1994). However, the simulation methods avoid calculation of the posterior normalizing constant, which is necessary for computing Bayes factors. Several ways of estimating the normalizing constant have been proposed (Carlin and Chib 1995; Chib 1995; Gelfand and Dey 1994; Gelman and Meng 1993; Green 1995; Kass and Wasserman 1992; Lewis and Raftery 1994; Meng and Wong 1993; Newton and Raftery 1994; Raftery 1995; Verdinelli and Wasserman 1995). It was not clear to us whether any of these is satisfactory for easy, routine use in a wide variety of well-behaved problems—well-behaved in the sense that the posterior would have a single, dominant mode. Thus we investigated several alternative methods, and modified some of these (in simple ways) to make them more effective. This article reports our results.

The Bayes factor for testing H_1 versus H_2 based on data y is

$$B_{12} = \frac{\int p_1(y|\beta)\pi_1(\beta)\,d\beta}{\int p_2(y|\theta)\pi_2(\theta)\,d\theta}, \tag{1}$$

where β and θ are parameters and π_1 and π_2 are the priors under the respective competing models p_1 and p_2. (For a review, see Kass and Raftery 1995.) The integrals in B_{12} have the form $C = \int h(\theta)d\theta$, where $h(\theta) = L(\theta)\pi(\theta)$ and $L(\theta) = p(y|\theta)$ with y fixed. We are concerned with the estimation of C via posterior simulation.

We consider a simulated version of Laplace's method, simulated versions of Bartlett correction, importance sampling, and a reciprocal importance sampling technique. We also introduce modifications to each of these. In addition, we apply the bridge sampling method of Meng and Wong (1993).

The modifications of Laplace's method and the Bartlett correction simply estimate the (unknown) value of the posterior probability density at the mode using the (simulated) probability assigned to a small region around the mode divided by its area. The importance sampling techniques are modified by restricting them to small regions about the mode. In Section 2 we describe the various methods for estimating C and their volume-corrected versions. In Section 3 we briefly discuss the problem of estimating the location and scale for the normal approximation to the posterior, which is used throughout. In Section 4 we provide theoretical remarks about these methods; in Section 5 we present numerical comparisons in simplified settings; and in Section 6 we analyze a nonlinear regression example. We present some closing remarks in Section 7. The main conclusion of this article is that the volume-corrected version of Laplace's method furnishes an accurate approximation that is especially useful when likelihood function evaluations are costly. A simple bridge-sampling technique used in conjunction with Laplace's method can achieve an order-of-magnitude improvement in accuracy.

2. METHODS FOR ESTIMATING C

Throughout, we assume that we have a sample $\theta_1, \ldots, \theta_m$ from the posterior distribution. For simplicity, when computing error rates, we assume that the sample is indepen-

Thomas DiCiccio is Associate Professor, Department of Social Statistics, Cornell University, Ithaca, NY 14853. Robert Kass is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Adrian Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Seattle, WA 98195. Larry Wasserman is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Kass and Wasserman's research was supported by National Institutes of Health grant RO1-CA54852 and National Science Foundation grant DMS-9303557, Wasserman's research was also supported by National Science Foundation- grant DMS-9357646, and Raftery's research was supported by Office of Naval Research grant N-00014-91-J-1074. The authors thank Xiao-Li Meng, two referees, and the associate editor for helpful comments. They also thank Alan Genz for providing the results of his methods applied to the example in Section 6.2.

dent, although in practice our methods can be used with dependent samples as well. We also assume that it is possible to evaluate the nonnormalized posterior $h(\theta) = L(\theta)h(\theta)$. Recall that the posterior density is $p(\theta|y) = h(\theta)/C$, where $C = \int h(\theta)d\theta$.

Note that $C = h(\theta)/p(\theta|y)$. If we can obtain an estimate of the posterior density at any point θ_0, then we can estimate C by $C = h(\theta_0)/p(\theta_0|y)$. This idea has been mentioned by Raftery (1995) and discussed in detail by Chib (1995). Laplace's method is a version of this; it uses a normal estimate of the posterior density. We discuss this in more detail in Section 2.6, but we take this opportunity to point out that we are interested mainly in problems in which estimating the posterior density is difficult, so that this strategy might not be feasible.

We denote the estimators of C based on Laplace's method, Bartlett correction, importance sampling, and reciprocal importance sampling by $\hat{C}_{\mathrm{L}}$, $\hat{C}_{\mathrm{B}}$, $\hat{C}_{\mathrm{I}}$, and $\hat{C}_{\mathrm{R}}$. In addition, we introduce alternatives to these that we denote by $\hat{C}^*_{\mathrm{L}}$, $\hat{C}^*_{\mathrm{B}}$, $\hat{C}^*_{\mathrm{I}}$, and $\hat{C}^*_{\mathrm{R}}$. We refer to the alternatives as "volume-corrected" or "localized." In this section we state the forms of the estimators. We defer theoretical remarks about $\hat{C}^*_{\mathrm{L}}$ and the Bartlett correction estimators to Section 4.

Let $\hat{\theta}$ be the posterior mode and let $\hat{\Sigma}$ be minus the inverse of the Hessian of the log-posterior evaluated at $\hat{\theta}$. If these cannot be obtained analytically, then they can be estimated via simulation; see Section 3. The normal approximation to the posterior is $\phi(\cdot) \equiv \phi(\cdot; \hat{\theta}, \hat{\Sigma})$, where $\phi(\cdot; \mathbf{a}, \mathbf{V})$ denotes a normal density with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{V}$. Let $B = \{\theta \in \Theta; \|(\theta - \hat{\theta})'\hat{\Sigma}^{-1}(\theta - \hat{\theta})\|^2 < \delta^2\}$, which has volume $v = \delta^p \pi^{p/2} |\hat{\Sigma}^{1/2}| / \Gamma(p/2 + 1)$. Let $P(B) = \int_B p(\theta|y)d\theta$ and let $\hat{P}$ be the Monte Carlo estimate of $P(B)$, that is, the proportion of the sampled values inside B. Also, we write $\Phi(B) \equiv \int_B \phi(\theta; \hat{\theta}, \hat{\Sigma})d\theta \equiv \alpha$.

2.1 Laplace Approximation

The Laplace approximation to C is obtained by approximating the posterior with a normal distribution. This approximation has a long history. (A good reference, which renewed interest in the method, is Tierney and Kadane 1986.) The approximation is given by

$$\hat{C}_{\mathrm{L}} = \frac{h(\hat{\theta})}{\phi(\hat{\theta}; \hat{\theta}, \hat{\Sigma})} = (2\pi)^{p/2} |\hat{\Sigma}|^{1/2} h(\hat{\theta}). \tag{2}$$

This approximation has error of order $O(n^{-1})$; that is, $C = \hat{C}_{\mathrm{L}}(1 + O(n^{-1}))$; see (A.2) in the Appendix.

The modification of $\hat{C}_{\mathrm{L}}$ is motivated by the observation that

$$C = \frac{h(\hat{\theta})}{p(\hat{\theta}|y)} = \frac{h(\hat{\theta})}{\phi(\hat{\theta})} \frac{\phi(\hat{\theta})}{p(\hat{\theta}|y)} \approx \frac{h(\hat{\theta})}{\phi(\hat{\theta})} \frac{\alpha}{P(B)}.$$

This observation suggests the volume-corrected estimator

$$\hat{C}^*_{\mathrm{L}} = \frac{h(\hat{\theta})}{\phi(\hat{\theta})} \frac{\alpha}{\hat{P}}, \tag{3}$$

which was discussed by Kass and Wasserman (1992).

2.2 Bartlett Adjustment

Bartlett adjustments are corrections that can improve first-order approximations. (See DiCiccio and Stern 1993 for a detailed discussion and further references.) The Bartlett-adjusted Laplace estimator is

$$\hat{C}_{\mathrm{B}} = \hat{C}_{\mathrm{L}} \cdot \left\{ \frac{\hat{E}(W|y)}{p} \right\}^{p/2}, \tag{4}$$

where $W(\theta) = 2\log(h(\hat{\theta})/h(\theta))$ and

$$\hat{E}(W|y) = \frac{1}{m} \sum_{i=1}^{m} W(\theta_i).$$

Ignoring simulation error, this approximation has error of order $O(n^{-2})$, and thus improves the Laplace estimator by an order of magnitude.

The local volume-corrected modification is defined by

$$\hat{C}^*_{\mathrm{B}} = \hat{C}^*_{\mathrm{L}} \cdot \left\{ 1 + \frac{\hat{E}(W|B, y) - N}{p + 2 - N} \right\}, \tag{5}$$

where

$$N = \frac{p}{\alpha} P(\chi^2_{p+2} \le \chi^2_p(\alpha)),$$

$$\hat{E}(W|B, y) = \frac{\sum_{i=1}^m W(\theta_i) Z_B(\theta_i)}{\sum_{i=1}^m Z_B(\theta_i)},$$

and Z_B is the indicator function for B.

2.3 Importance Sampling

Importance sampling is a common technique for estimating expectations using simulation (Geweke 1989; Hammersley and Handscomb 1964). The constant C may be estimated by importance sampling as follows. Draw $\tilde{\theta}_1, \ldots, \tilde{\theta}_M$ from a distribution Q with density q. Then

$$\hat{C}_{\mathrm{I}} = \frac{1}{M} \sum_i \frac{h(\tilde{\theta}_i)}{q(\tilde{\theta}_i)}. \tag{6}$$

An important practical problem is the choice of q. Generally, to reduce the variance of the ratio $h(\tilde{\theta}_i)/q(\tilde{\theta}_i)$, q should be similar to and have tails no thinner than h. In this article we use the sample from the posterior to choose q. In particular, we take $q(\cdot) = \phi(\cdot; \hat{\theta}, \hat{\Sigma})$. This is slightly different from the usual importance sampling method, because we are using a sample from the posterior to choose the importance sampling function. Our approach is to consider only this simple choice and see how it and its locally restricted version behave.

The locally restricted version is based on the identity

$$C = \frac{1}{P(B)} E_Q \left(\frac{h(\theta) Z_B(\theta)}{q(\theta)} \right) = \frac{Q(B)}{P(B)} E_{Q_B} \left(\frac{h(\theta)}{q(\theta)} \right), \tag{7}$$

where Z_B is the indicator function for B and $Q_B(\cdot) = Q(\cdot|B)$. The first factor on the right side of (7) can be estimated from the original sample. The second factor can

be estimated using the sample from Q. This gives the local importance sampling estimate

$$\hat{C}_{\mathrm{I}}^{*} = \frac{\frac{1}{M}\sum_i h(\tilde{\theta}_i) Z_B(\tilde{\theta}_i)/q(\tilde{\theta}_i)}{\frac{1}{m}\sum_i Z_B(\theta_i)}. \tag{8}$$

Alternatively, it is usually more efficient to sample from Q_B and use the expression on the right side of (7) directly.

2.4 Reciprocal Importance Sampling

The disadvantage of importance sampling is that it requires a second sample from Q. It is possible to use the posterior sample directly. Gelfand and Dey (1994) and Tierney (personal communication) proposed the estimator

$$\hat{C}_{\mathrm{R}} = \left\{ \frac{1}{m} \sum_i \frac{s(\theta_i)}{h(\theta_i)} \right\}^{-1}, \tag{9}$$

where $s(\cdot)$ is an arbitrary probability density function. Choosing $s = \pi$ gives the harmonic mean estimator proposed by Newton and Raftery (1994). However, taking $s = \pi$ leads to an estimator of C^{-1} that may have infinite variance. If s has tails that are thin enough—specifically, if $\int s^2/h < \infty$—then this estimator will be well behaved (Newton and Raftery 1994, p. 47). This is the opposite of importance sampling, in which the importance sampling function must have thicker tails than the target distribution. Like importance sampling, s must be similar to h for $\hat{C}_{\mathrm{R}}$ to have a small variance. In practice, it may be hard to find a density s that has sufficiently thin tails in all directions of the parameter space simultaneously, especially when the posterior has nonelliptical contours. This is discussed further in Section 5.

Our modified version is obtained by restricting calculations to the ellipse B and by taking $s(\cdot) = \phi(\cdot; \hat{\theta}, \hat{\Sigma})$ in (9). Then define the local reciprocal importance sampling estimate

$$\hat{C}_{\mathrm{R}}^{*} = \alpha \left\{ \frac{1}{m} \sum_i \frac{s(\theta_i) Z_B(\theta_i)}{h(\theta_i)} \right\}^{-1}. \tag{10}$$

The hope is that by restricting the sum to B, we can avoid instances where s/h is large. Note that if we make the approximation $s(\theta)/h(\theta) \approx s(\hat{\theta})/h(\hat{\theta})$ over B, then $\hat{C}_{\mathrm{R}}^{*} \approx \hat{C}_{\mathrm{L}}^{*}$. To have finite variance, we need $\int_B s^2/h < \infty$, which is clearly easier to achieve than $\int s^2/h < \infty$. Thus the adjusted version may be more stable.

2.5 Bridge Sampling

The foregoing two methods are special cases of bridge sampling, a class of techniques analyzed by Meng and Wong (1993) for estimating the ratio of two normalizing constants; the method dates back to Bennett (1976). This technique arises from the following identity. Let $s_1 = t_1/c_1$ and $s_2 = t_2/c_2$ be two densities where $c_i = \int t_i$, for $i = 1, 2$. For our purposes, we assume that both densities have the same support. Let γ be a function satisfying $0 < |\int \gamma(\theta) s_1(\theta) s_2(\theta) d\theta| < \infty$. Then we have that

$$\frac{c_1}{c_2} = \frac{\int t_1(\theta)\gamma(\theta) s_2(\theta)\, d\theta}{\int t_2(\theta)\gamma(\theta) s_1(\theta)\, d\theta}. \tag{11}$$

If we let $t_1 = h$, $c_1 = C$, $t_2 = q$, and $c_2 = 1$, where q is the normal approximation to the posterior or any other convenient approximation, then the identity becomes

$$C = \frac{\int h(\theta)\gamma(\theta) q(\theta)\, d\theta}{\int q(\theta)\gamma(\theta) p(\theta|y)\, d\theta}. \tag{12}$$

Now draw a sample $\tilde{\theta}_1, \ldots, \tilde{\theta}_M$ from q. Recall that we also have a sample $\theta_1, \ldots, \theta_m$ from the posterior. Then the Meng–Wong bridge estimator becomes

$$\hat{C}_{\mathrm{MW}} = \frac{\frac{1}{M}\sum_i h(\tilde{\theta}_i)\gamma(\tilde{\theta}_i)}{\frac{1}{m}\sum_i q(\theta_i)\gamma(\theta_i)}. \tag{13}$$

In terms of a mean squared error (MSE) criterion, Meng and Wong showed that the optimal choice of γ for a given q is proportional to $\{mh(\theta)/C + Mq(\theta)\}^{-1}$. This involves knowing C, but they also discussed iterative methods for finding γ. Choosing $\gamma = q^{-1}$ reduces the method to ordinary importance sampling with importance sampling function q, and choosing $\gamma = Z_B/q$ gives local importance sampling. Choosing $\gamma = h^{-1}$ reduces the method to reciprocal importance sampling, whereas $\gamma = Z_B/h$ gives local reciprocal importance sampling. In our implementation here, we take q to be the normal approximation to the posterior density. Thus $\hat{C}_{\mathrm{I}}^{*}$ of Section 2.3 becomes a suboptimal bridge sampling method in the sense of Meng and Wong. (This was pointed out to us in a personal communication from X.-L. Meng and W. Wong.)

2.6 Other Methods

Verdinelli and Wasserman (1995) suggested an estimator that is useful when the two models being tested are nested. Suppose that the model is $p(y|\omega, \psi)$ and that we wish to test H_1: $\omega = \omega_0$ versus H_2: $\omega \neq \omega_0$. If $\pi_2(\psi|\omega_0) = \pi_1(\psi)$, then $B_{12} = p_2(\omega_0|y)/\pi_2(\omega_0)$ (the "Savage–Dickey density ratio"), and hence estimating B_{12} comes down to estimating $p_2(\omega_0|y)$, which can be done by ordinary density estimation methods. If $\pi_2(\psi|\omega_0) \neq \pi_1(\psi)$, then Verdinelli and Wasserman (1995) noted that the following identity holds:

$$B_{12} = p_2(\omega_0|y) E_{\psi|\omega_0, y}\left(\frac{\pi_1(\psi)}{\pi_2(\omega_0, \psi)} \right),$$

where the expectation is with respect to the posterior under H_2 with ω fixed at ω_0. As before, the first term can be estimated by density estimation techniques. The second term can be estimated by simulation from the posterior distribution under H_1; thus a second simulation may be required. This method is very effective when the dimension of ω is not too large.

Another method for estimating C is to draw a sample $\theta_1, \ldots, \theta_m$ from the prior π and then take $\hat{C} = m^{-1}\sum_i p(y|\theta_i)$. But this approach is not efficient, because the likelihood is usually very concentrated relative to the prior, and hence most sampled points fall into regions where the likelihood is small.

As mentioned earlier, the simple identity, $C = h(\theta)/p(\theta|y)$ for all θ, suggests estimating C simply by inserting an estimate of $p(\theta|y)$ at one point. Raftery (1995) called the resulting estimator the "candidate's estimator." One particularly simple kernel density estimator is just the number of points in B divided by its volume, which yields the easily implemented estimator $\hat{C}_{\mathrm{C}} = h(\hat{\theta})v/\hat{P}$. This estimator is valid for any definition of $\hat{\theta}$ (not just the posterior mode) as long as the ellipse B is centered at $\hat{\theta}$, and it is likely to be reasonably efficient as long as $\hat{\theta}$ is fairly close to the posterior mode. Further discussion of $\hat{C}_{\mathrm{C}}$ is presented in Section 4.1.

In some cases better estimates of $p(\theta|y)$ may be found, particularly when convenient latent variables are available (see Chib 1995). As Chib noted, $p(\theta|y)$ can be estimated by a sequence of simulations followed by a sequence of one-dimensional (or lower-dimensional) density estimates. For example, if we write $\theta = (\theta^1, \ldots, \theta^p)$, then

$$p(\theta|y) = p(\theta^1|y)p(\theta^2|\theta^1, y)\ldots p(\theta^p|\theta^1, \ldots, \theta^{p-1}, y).$$

The jth term $p(\theta^j|\theta^1, \ldots, \theta^{j-1}, y)$ can be estimated by kernel density estimation using a simulation from $p(\theta^j|\theta^1, \ldots, \theta^{j-1}, y)$. There is a trade-off here: Choosing each θ^i to be of low dimension makes the density estimation easier but increases the required number of simulations. We illustrate this method in Section 6.2.

Carlin and Chib (1995) showed how to use Gibbs sampling to estimate Bayes factors. They included an indicator function for the true model as a parameter. To obtain a well-defined Gibbs sampler, one needs the conditional distribution of every parameter given all of the others. In particular, one needs the conditional distribution of the parameters of one model given that the other model is the true model. This generally is not a well-defined object, so Carlin and Chib created "linking densities" to produce a well-defined Gibbs sampler.

Green (1995) has argued that the Carlin–Chib approach may be cumbersome and inefficient because of the necessity of creating and sampling from the linking densities. Green proposed a more direct method, based on Metropolis sampling, in which a Markov chain is used to move around within models and between models in such a way that the limiting proportion of visits to a given model is the posterior probability of that model. Suppose that model 1 has one parameter, ψ, and model 2 has two parameters, θ_1 and θ_2. One must construct a Markov chain that moves around within model 1 and within model 2 and that sometimes jumps between the two models. For example, in jumping from (θ_1, θ_2) in model 2 to model 1, we might generate a random draw from a distribution centered at $\psi = (\theta_1 + \theta_2)/2$. Finding sensible, efficient ways to jump between the models requires some insight and must be approached on a problem-by-problem basis.

Gelman and Meng (1994) proposed a generalization of bridge sampling that they called path sampling. This involves constructing a continuous path of densities between p and q. We do not pursue path sampling further in this article, although we note that the method shows some promise. Phillips and Smith (1994) used jump diffusions to estimate normalizing constants.

Our focus in this article is slightly different than some of the aforementioned methods in that we emphasize methods that use the output of the simulation from a given posterior, possibly followed by one simple extra simulation.

3. ESTIMATING THE LOCATION AND SCALE OF THE POSTERIOR BY SIMULATION

All of the methods require some estimate of the location and scale of the posterior. In some cases these are available analytically or by standard numerical methods; for example, one might use Newton–Raphson or Fisher's method of scoring. However, it is much simpler if the output of the simulation from the posterior can be used directly to estimate location and scale, as noted by Kass and Wasserman (1992), Lewis and Raftery (1994), and Raftery (1995). This is especially true in high-dimensional problems where performing an optimization in addition to a simulation may be a burden. In this section we briefly describe methods of using simulation output to estimate the location and scale.

In practice, a simple, crude method will usually suffice. A naive approach is to use $\hat{\theta} \approx \bar{\theta}$ or $\hat{\theta} \approx \mathrm{argmax}_i h(\theta_i)$ where $\bar{\theta}$ is the sample average of the simulated values. Similarly, we may use $\hat{\mathbf{\Sigma}} \approx \bar{\mathbf{\Sigma}}$, the sample covariance matrix. Our experience is that these can be poor estimates, however.

One way of improving these estimates is to select the points that fall in the small ellipse B around the starting estimate of the mode. Let $l(\theta) = \log h(\theta)$. Because the log posterior is approximately quadratic, we fit a second-order regression to the selected points, say, $l(\theta) = b_0 + b'\theta + \theta' G\theta$, where $\theta' = (\theta_1, \ldots, \theta_p)$, $b' = (b_1, \ldots, b_p)$, $G = \{g_{ij}\}$, $g_{ii} = b_{ii}$, and $g_{ij} = g_{ji} = (1/2)b_{ij}$ for $i \neq j$. Setting the derivative equal to 0 gives an improved estimate of the mode, namely $-(1/2)G^{-1}b$. Similarly, an estimate of $h(\hat{\theta})$ is $\exp\{b_0 - (1/4)b'G^{-1}b\}$, and an estimate of $\hat{\Sigma}$ is $-(1/2)G^{-1}$.

Lewis and Raftery (1994) and Raftery (1995) suggested estimating $\hat{\theta}$ by finding the value of θ_i that minimizes $\sum_j |\theta_i - \theta_j|$ or by using the componentwise median. They also suggested estimating $\hat{\Sigma}$ by the minimum volume ellipsoid method of Rousseeuw and van Zomeren (1990), which also provides a robust estimate of location. Our experience is that the median and median absolute deviation work well in one dimension. We also found the componentwise median to work well as an estimator of $\hat{\theta}$ in higher dimensions. The Rousseeuw and van Zomeren method works well in higher dimensions but can be very slow. Other robust methods have been discussed by Gnanadesikan (1976, sec. 5.2.3).

4. THEORETICAL REMARKS

In this section we discuss the theory behind some of the estimators presented in Section 2.

4.1 Laplace's Method

As we remarked in Section 2.1, $C = \hat{C}_{\mathrm{L}}(1 + O(n^{-1}))$. The local version of $\hat{C}_{\mathrm{L}}$ is essentially $\hat{C}_{\mathrm{L}}$ times a

"histogram-like" correction factor. In fact, one can estimate C directly by using a local histogram around an arbitrary point θ, which yields the simple "candidate's estimator," $\hat{C}_{\mathrm{C}}$, discussed in Section 2.6. To motivate the local Laplace estimator, we first discuss this local histogram estimator. Let n be fixed, let m tend to infinity, and let $\delta = \delta_m$ be a function of m. Recall that B has volume $v = \delta^p \pi^{p/2} |\hat{\Sigma}^{1/2}| / \Gamma(p/2+1)$, so $v = O(\delta^p)$, because n is fixed. The approximation $\hat{C}_{\mathrm{C}}$ is based on the idea of approximating the posterior density at its mode by computing the probability of a small set B containing the mode and dividing by the volume of B. The approximation $\hat{C}^*_{\mathrm{L}}$ improves on this by using the fact that the posterior is nearly normal over B.

We define $\hat{C}_{\mathrm{C}}$ as follows:

$$C = \frac{h(\hat{\theta})}{p(\hat{\theta}|y)} \approx \frac{vh(\hat{\theta})}{P(B)} \approx \frac{vh(\hat{\theta})}{\hat{P}} \equiv \hat{C}_{\mathrm{C}}.$$

Let $P = P(B_\delta)$. Recall that δ and hence P are functions of m, but for simplicity we suppress this dependence. Now by expanding the posterior density, we have $P = vp(\hat{\theta}|y) + vO(\delta^2)$ and $P - \hat{P} = O(\{\delta^p/m\}^{1/2})$, so that $P/\hat{P} = 1/(1 + O((\delta^p m)^{-1/2}))$. Hence

$$\begin{aligned}\hat{C}_{\mathrm{C}} &= \frac{vh(\hat{\theta})}{\hat{P}} = \frac{h(\hat{\theta})}{P/v}\,\frac{P}{\hat{P}} \\ &= \frac{h(\hat{\theta})}{p(\hat{\theta}|y) + O(\delta^2)}\,\frac{1}{(1 + O((\delta^p m)^{-1/2}))} \\ &= C(1 + O(\delta^2))(1 + O((\delta^p m)^{-1/2})).\end{aligned}$$

There are two sources of error: the first from approximating a function by its average over B, and the second from Monte Carlo error. The error tends to 0 as m goes to infinity provided that we choose a sequence δ_m such that $\delta_m \to 0$ and $\delta_m^p m \to \infty$. The best achievable error rate is attained by taking $\delta_m = O(m^{-1/(4+p)})$, for which $\hat{C}_{\mathrm{C}} = C(1 + O(m^{-2/(4+p)}))$. In principle, this error can be made as small as we like by making m large. The problem of choosing a good value of δ is the standard problem of choosing window width or bandwidth in density estimation.

We seek to improve $\hat{C}_{\mathrm{C}}$ by taking advantage of the fact that the posterior becomes normal as n tends to infinity. Let n increase, let $m = m_n$ be a function of n, and note that δ is also a function of n. Now

$$C = \frac{h(\hat{\theta})}{\phi(\hat{\theta})}\,\frac{\phi(\hat{\theta})}{p(\hat{\theta}|y)} \approx \frac{h(\hat{\theta})}{\phi(\hat{\theta})}\,\frac{\Phi(B)}{P(B)},$$

and, recalling that $\alpha = \Phi(B)$, we define

$$\hat{C}^*_{\mathrm{L}} = \frac{h(\hat{\theta})}{\phi(\hat{\theta})}\,\frac{\alpha}{\hat{P}}.$$

By expanding C in a fourth-order series about $\hat{\theta}$, we have that

$$C = \frac{h(\hat{\theta})}{\phi(\hat{\theta})}\,\frac{\Phi(B)}{P(B)}\,(1 + O(n^{-1}\delta^4)).$$

If we multiply and divide by $\hat{P}$ and use the fact that $P/\hat{P} = 1/(1 + O((\delta^p m)^{-1/2}))$, we get

$$\hat{C}^*_{\mathrm{L}} = C \cdot (1 + O(n\delta^4))(1 + O((m\delta^p)^{-1/2})).$$

It is apparent that the Monte Carlo errors in $\hat{C}_{\mathrm{C}}$ and $\hat{C}^*_{\mathrm{L}}$ are of the same order, so that when these are neglected, $\hat{C}^*_{\mathrm{L}}$ becomes more accurate than $\hat{C}_{\mathrm{C}}$ when $\delta = o(1)$. However, if we treat this as a density estimation problem and let the size of the ellipse shrink at rate $m^{-1/(4+p)}$, then the modification will improve on the Laplace rate if $mn^{-(4+p)/2} \to \infty$. This simply means that the modification works if the simulation size is large relative to the sample size. The modified Laplace estimator is very similar to the semiparametric density estimator recently proposed by Hjort and Glad (1995).

There is a bias–variance trade-off in choosing α. To see this, note that

$$E\left(\frac{C}{\hat{C}^*_{\mathrm{L}}} - 1\right)^2 = A^2 \frac{1 - P(B_\delta)}{P(B_\delta)m} + (A-1)^2,$$

where

$$A = \frac{\phi(\hat{\theta})P(B_\delta)C}{h(\hat{\theta})Q(B_\delta)}.$$

The first term goes to 0 as $\delta \to \infty$, and the second term goes to 0 as $\delta \to 0$. Hence small ellipses have small bias and large variance, and large ellipses have large bias and small variance.

4.2 Bartlett Adjustments

The details of the Bartlett adjustments are in the Appendix. Here we outline the main results. For simplicity, it is convenient to derive the Bartlett-adjusted Laplace estimator and its local version in the one-dimensional case $p = 1$. Let $H(\theta) = \log h(\theta)$, $H_j(\theta) = d^j H(\theta)/d\theta^j$, and $\hat{H}_j = H_j(\hat{\theta})$, $(j = 1, \ldots, p)$; hence $-\hat{H}_2 = \hat{\Sigma}^{-1}$. By including further terms in the Taylor approximation, it can be shown that

$$\begin{aligned}C &= h(\hat{\theta})(2\pi)^{1/2}\hat{\Sigma}^{1/2}\{1 + \kappa/2 + O(n^{-2})\} \\ &= \hat{C}_{\mathrm{L}}\{1 + \kappa/2 + O(n^{-2})\},\end{aligned}$$

where

$$\kappa = \frac{1}{4}\,\hat{H}_4\hat{\Sigma}^2 + \frac{5}{12}\,\hat{H}_3^2\hat{\Sigma}^3 = O(n^{-1}).$$

The quantity κ also arises in an asymptotic formula for the posterior expectation of the posterior ratio statistic $W = 2\{H(\hat{\theta}) - H(\theta)\}$. To error of order $O(n^{-2})$, $E(W|y) = 1 + \kappa + O(n^{-2})$ and

$$\begin{aligned}C &= \hat{C}_{\mathrm{L}}\left\{1 + \frac{E(W|y) - 1}{2} + O(n^{-2})\right\} \\ &= \hat{C}_{\mathrm{L}}[\{E(W|y)\}^{1/2} + O(n^{-2})].\end{aligned}$$

In the vector case we get

$$C = \hat{C}_{\mathrm{L}} \left\{ 1 + \frac{E(W|y) - p}{2} + O(n^{-2}) \right\}$$

$$= \hat{C}_{\mathrm{L}} \left[\left\{ \frac{E(W|y)}{p} \right\}^{p/2} + O(n^{-2}) \right].$$

The second approximation motivates the Bartlett-adjusted Laplace estimator. These formulas were considered by DiCiccio and Stern (1993). The term $(1 + \kappa/p)^{-1}$ is a Bayesian Bartlett adjustment. DiCiccio and Stern (1993) showed that the posterior distribution of the Bartlett-adjusted posterior ratio statistic $(1 + \kappa/p)^{-1}W$ is chi-squared to error of order $O(n^{-2})$.

It is plausible that the accuracy of asymptotic approximations to C and $E(W|y)$ could be improved in practice by taking the interval of integration $I_k = [\hat{\theta} - k, \hat{\theta} + k]$ into account explicitly. If the integrals under study are actually restricted to that interval and at least the leading terms of the expansions are adjusted accordingly, then much of the error induced by the processes of first truncating the range of integration and later enlarging it to encompass the whole real line could be largely eliminated. The remaining errors would, in principle, arise primarily from the Taylor approximations to the integrands.

Let $P_k = P(I_k)$ be the posterior probability content of the interval I_k, and let $\Phi_k = \Phi(I_k)$. Taking the interval I_k into account leads to different formulas. Let $N_k = E\{(\theta - \hat{\theta})^2 \hat{\Sigma}^{-1} | \theta \in I_k\}$ with $\theta \sim N(\hat{\theta}, \hat{\Sigma})$. Then

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{E(W|I_k, y) - N_k}{3 - N_k} + O(n^{-2}) \right\}.$$

A version of this expression is available in the vector case as outlined in the Appendix. The final formula is the same, except that N_k takes its obvious vector version and $3 - N_k$ is replaced by $p + 2 - N_k$, with $B_k = \{\theta \in \Theta | (\theta - \hat{\theta})' \hat{\Sigma}^{-1} (\theta - \hat{\theta}) \le nk^2\}$; the factor n is present to balance the introduction of $\hat{\Sigma}^{-1}$, whose elements are of order $O(n)$. Note that the volume of B_k is of order $O(1)$, and thus B_k generalizes the fixed-length interval I_k considered in the one-dimensional case. Note that $N_k = p(\Pr\{\chi^2_{p+2} \le nk^2\} / \Pr\{\chi^2_p \le nk^2\})$.

A key assumption in the preceding calculations is that the regions of integration are nonshrinking; recall that k was understood to be constant. In calculations (A.10) and (A.12) of the Appendix, this assumption justified replacing the interval I_k by the whole real line to obtain the terms of order $O(n^{-1})$. The calculations remain valid even if the region of integration is shrinking, provided that the rate of shrinkage is sufficiently slow. In particular, the conclusions hold if $n^{1/2}k \to \infty$ and if for such k, the probability contents P_k and Φ_k both tend to 1. For practical implementation, a region $B_\delta = \{\theta \in \Theta; (\theta - \hat{\theta})' \hat{\Sigma}^{-1} (\theta - \hat{\theta}) \le \delta^2\}$ must be specified, where δ^2 corresponds to nk^2. The present asymptotic approach indicates choosing δ so that $\alpha = \Pr\{\chi^2_p \le \delta^2\}$ is moderately large, say between .5 and .8.

Another, although arguably less intuitive, viewpoint is to regard δ as fixed; that is, a probability α is chosen independently of n, with the integration restricted to regions having normal probability content α. In this framework the quantity k that determines B_k is of order $O(n^{-1/2})$. When k is of order $O(n^{-1/2})$, the region of integration must also be taken into account for calculating higher-order terms. In the one-dimensional case, we find that

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{E(W|I_k, y) + D_k - N_k}{3 - N_k} + O(n^{-2}) \right\},$$

where D_k and κ' are defined in the Appendix.

The formula for C contains terms other than N_k and κ'. Because of these extra terms, it appears that in the case where δ is fixed and k is of order $O(n^{-1/2})$, the Bartlett adjustment method for improving the local Laplace approximation is not strictly valid. Curiously, and somewhat counterintuitively, if k shrinks much faster than $O(n^{-1/2})$, say if $k = o(n^{-3/4})$, then the remaining terms in the last expression are $O(n^{-2})$, so again the Bartlett adjustment is apparently valid. We do not pursue this curiosity further here. However, it raises the interesting possibility that Bartlett adjustments may be effective for very small α. Preliminary numerical investigations (not reported here) confirm this phenomenon.

5. EXAMPLES BASED ON THE SKEWED-NORMAL AND SKEWED-T DISTRIBUTIONS

It is well known that importance sampling can be inefficient when the importance sampling function q has thin tails relative to h (see, e.g., Geweke 1989). Reciprocal importance sampling turns out to be inefficient when h is skewed and thin-tailed. To see this, suppose that h is skewed and s is a normal density. Then there might be regions where h is small relative to s, causing $\hat{C}_{\mathrm{R}}$ to blow up. Similar problems occur in multiparameter problems when the posterior has banana-shaped contours. This is not at all a pathological phenomenon; indeed, it occurs often in practice.

To permit closer examination of the methods in light of this observation, we consider examples where the underlying density is skewed. A convenient family of densities for this purpose is the skewed t distribution, which includes the skewed normal (Azzalini 1985; O'Hagan and Leonard 1976). A random variable V has a (standard) skewed normal distribution, denoted by $V \sim \mathrm{SN}(\lambda)$, if it has density $f(z|\lambda) = 2\phi(z)\Phi(\lambda z)$. Here ϕ is the standard normal density, Φ is the standard normal cdf and λ is the skewness parameter. It is simple to see (although we have not seen this mentioned before), that V has the following latent variable interpretation: draw $W \sim N(0, 1)$; with probability $\Phi(\lambda W)$, set $V = W$, and with probability $1 - \Phi(\lambda W)$, set $V = -W$. Thus a skewed normal random variable is just a standard normal random variable with a random sign change. This latent variable structure makes it simple to simulate from a skewed normal distribution. If we let W have a t distribution with ν degrees of freedom instead of a normal, then we say that V has a skewed t_ν distribution and write $V \sim \mathrm{ST}(\lambda, \nu)$. Of course, a $\mathrm{ST}(\lambda, \infty)$ distribution is equivalent to a $\mathrm{SN}(\lambda)$ distribution. Denote the skewed-t density by $f(z|\lambda, \nu)$.

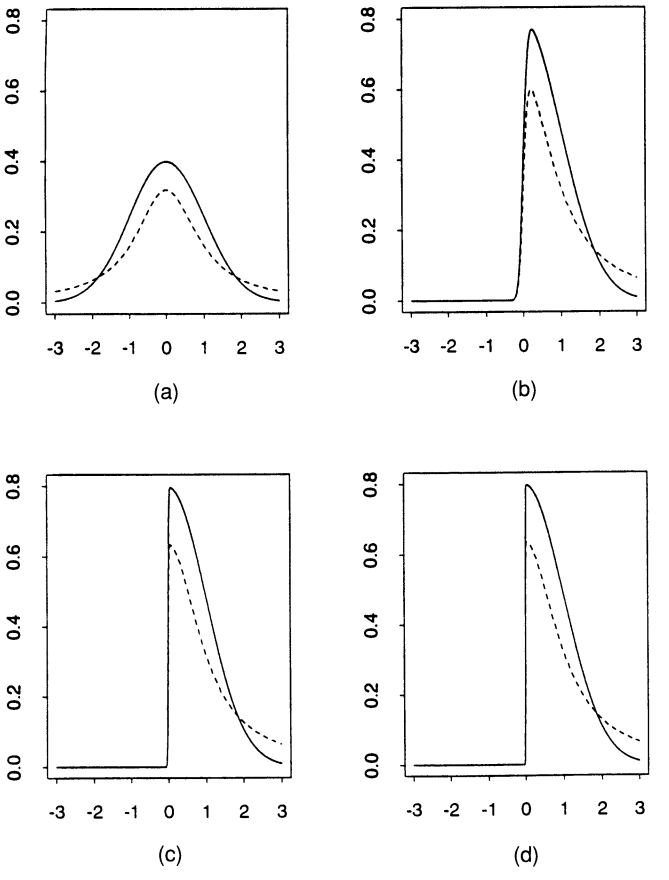

Figure 1. The Skewed Normal (Solid Lines) and Skewed Cauchy (Dashed Lines) for λ = 0 (a), 10 (b), 50 (c), and 100 (d). When λ = 100, the densities are so extremely skewed that they are close to being a half-normal and half-Cauchy.

We let $h(z) = f(z|\lambda, \nu)$, so that $C = 1$, and we examined two cases: $\lambda = 100$ and $\nu = \infty$, and $\lambda = 100$ and $\nu = 1$. As shown in Figure 1, these densities are very highly skewed—much more skewed, in fact, than likely would be found in practice.

For these two densities we computed the Laplace, Bartlett correction, importance sampling, and reciprocal importance sampling estimators and their modifications. We also computed the Meng–Wong bridge estimator. Recall that for the bridge estimator estimator, one must choose the function γ, defined in Section 2.5. The optimal choice derived by Meng and Wong involves the constant C. We computed two versions for the bridge estimator: the optimal version (with C set to the true value in the formula for γ) and a "Laplace bridge estimator" with C set to $\hat{C}_{\mathrm{L}}$. We used m = 10,000, m = 100,000, $\alpha = .05$, and $\alpha = .50$. For bridge sampling, we took each of the two simulations to be of size m. We repeated the experiment 100 times for both distributions and computed $(1/100)\sum_j |\log \hat{C}_j|$ for each of the methods. (This is approximately the geometric mean relative error.) The results are reported in Table 1. We also carried out a simulation using the product of five densities (so that $p = 5$) for each of the two types of skewed t distributions described previously for Table 1. The five-dimensional results are not reported here, although these results are qualitatively similar to those for the one-dimensional case.

Several findings are apparent. First, given that the distributions are extremely skewed and that only rough accuracy is necessary for Bayes factor calculations (an error of magnitude .5 on the $\log_e$ scale would be quite tolerable in practice), all methods appear to be reasonably accurate. Second, in nearly all cases, the modifications substantially improved the first four methods. Restricting to neighborhoods of the mode can be very beneficial for importance sampling and reciprocal importance sampling. Third, the modified Laplace estimator $\hat{C}^*_{\mathrm{L}}$ is generally more accurate for $\alpha = .05$ than for $\alpha = .5$ (as predicted by the theory discussed in Sec. 4). Fourth, $\hat{C}^*_{\mathrm{L}}$ is generally quite accurate, but is sometimes a little less accurate than the modified Bartlett correction estimator $\hat{C}^*_{\mathrm{B}}$, and can be quite a bit less accurate than $\hat{C}^*_{\mathrm{R}}$ and, especially, $\hat{C}^*_{\mathrm{I}}$. Fifth, optimal bridge sampling based on the normal approximation to the posterior is very effective, increasing accuracy by a factor of 10 or more in some cases. This procedure may be motivated by the good results produced by $\hat{C}^*_{\mathrm{I}}$ together with the observation (made in Sec. 2.5) that $\hat{C}^*_{\mathrm{I}}$ is a bridge sampler, but it is suboptimal. Finally, and quite interestingly, the Laplace bridge sampling estimator is, for practical purposes, just as accurate as optimal bridge sampling.

We also investigated the relative mean squared error RMSE $= E(C/\hat{C}^*_{\mathrm{L}} - 1)^2$ as a function of α. Plots of this quantity by α (not included) reveal the following behavior. Except when the skewness is extreme, RMSE is a remarkably stable function of α. When extreme skewness is present, $\alpha = .05$ is optimal or near optimal for m between 1,000 and 100,000. Generally, the penalty for choosing α too small is greater than the penalty for choosing α too large. The results suggest that $\alpha = .05$ is a reasonable default value.

6. FURTHER EXAMPLES

In this section we consider two nonlinear regression examples. The first is two-dimensional but has an extremely nonnormal posterior. The second is 10-dimensional and is included to explore the feasibility of the methods in higher dimensions.

6.1 A Two-Dimensional Nonlinear Regression

We consider data on biochemical oxygen demand (BOD) collected by D. Marske at the University of Wisconsin-Madison and described by Bates and Watts (1988, p. 270). The model used by Bates and Watts is

$$Y_i = \theta_1(1 - e^{-\theta_2 X_i}) + \varepsilon_i,$$

where the ε_i's are independent $N(0, \sigma^2)$ errors, Y_i is BOD (mg/L), and X_i is time (days). As shown by Bates and Watts (1988, p. 202), the likelihood contours are highly nonelliptical. The example thus provides an interesting testing ground for the various methods.

We take $p(\sigma) \propto \sigma^{-1}$ and integrate out σ. This leaves θ_1 and θ_2. For the sake of illustration, we take $\theta_1 \sim U(0, 60)$ and $\theta_2 \sim U(0, 6)$. The log-normalizing constant for the posterior was found by numerical integration to be $\log C = -16.205$. We then used the four methods and their modified

Table 1. Comparison of the Methods: The mean absolute value of log $\hat{C}$

	Laplace	*Bartlett*	*Reciprocal*	*Importance*	*Laplace bridge*	*Optimal bridge*
			ST (100, ∞) Density			
α = .05, m = 10,000						
Original	.060	.047	.124	.007	.004	.004
Modified	.037	.037	.037	.037	NA	NA
α = .50, m = 10,000						
Original	.060	.046	.124	.006	.005	.005
Modified	.059	.024	.008	.007	NA	NA
α = .05, m = 100,000						
Original	.060	.053	.120	.002	.001	.001
Modified	.012	.018	.012	.012	NA	NA
α = .50, m = 100,000						
Original	.060	.046	.123	.002	.001	.001
Modified	.060	.023	.002	.002	NA	NA
			ST (100, 1) Density			
α = .05, m = 10,000						
Original	.144	.366	.189	.110	.006	.005
Modified	.038	.056	.038	.038	NA	NA
α = .50, m = 10,000						
Original	.144	.367	.189	.120	.006	.006
Modified	.144	.106	.010	.010	NA	NA
α = .05, m = 100,000						
Original	.143	.368	.183	.106	.002	.002
Modified	.013	.040	.013	.013	NA	NA
α = .50, m = 100,000						
Original	.144	.367	.185	.113	.003	.003
Modified	.144	.107	.003	.003	NA	NA

NOTE: The true value of *log C* is 0.

versions based on a Metropolis sampling scheme (Tierney 1994) with 10,000 iterations. We repeated this process 10 times. The relative error of each method (averaged over the 10 independent replications) is shown in Table 2.

All of the methods, except the reciprocal importance sampling, were improved by using the modified version. However, the modified version of reciprocal importance sampling still does reasonably well. Localization tends to improve reciprocal importance sampling when the posterior is thin-tailed. In this example the likelihood tends to die off slowly in some directions. The modified Laplace method does reasonably well. As in the simulations, the Laplace bridge estimator does best, producing an estimate with less than 10% error.

Table 2. Relative Error $|\hat{C} - C|/C$ in Estimating the Normalizing Constant in the BOD Example

	Laplace	*Bartlett*	*Reciprocal*	*Importance*	*Laplace bridge*
Original	.181	.415	.064	.227	.070
	(.013)	(.027)	(.012)	(.008)	(.020)
Modified	.126	.232	.137	.138	NA
	(.022)	(.043)	(.025)	(.020)	NA

NOTE: The error is averaged over 10 replications. The estimates for each replication are based on a Metropolis sampling chain of length 10,000. The numbers in parentheses are the standard errors of the estimates of errors, based on the 10 replications.

6.2 A 10-Dimensional Example

Next, we consider a 10-dimensional nonlinear regression example. The data, from Bates and Watt (1988, p. 275), are on the kinematic viscosity of a lubricant as a function of temperature, x_1, and pressure, x_2. The proposed model is

$$E(Y|\theta) = \frac{\theta_1}{\theta_2 + x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3 + (\theta_6 + \theta_7 x_2^2) x_2 \exp\left\{ - \frac{x_1}{\theta_8 + \theta_9 x_2^2} \right\}$$

with $N(0, \sigma^2)$ errors. The tenth parameter is $\theta_{10} = \log \sigma$. (See Bates and Watts 1988, p. 89, for an analysis of these data.) For the purpose of illustration only, we adopt independent normal priors centered at the maximum likelihood estimators. We take the prior standard deviation for a parameter to be $n^{1/2}$ times the standard error of the maximum likelihood estimate of that parameter. This makes the prior have approximately the amount of information contained in one observation; Kass and Wasserman (1995) discussed priors based on one unit of information in testing problems. We emphasize that we are not necessarily recommending this prior for a substantive analysis of this problem.

We obtained our results using Markov chain Monte Carlo. The chains were run for 120,000 iterations, and the first 20,000 were discarded. At the suggestion of a referee, we also included a version of Chib's estimator. Specifically,

Table 3. Relative Error $|\hat{C} - C|/C$ in Estimating the Normalizing Constant in the Lubricant Example

	Laplace	Bartlett	Reciprocal	Importance	Laplace bridge	Bridge (5)	Chib
Original	.613	.637	.588	.095	.486	.469	.427
	(.107)	(.052)	(.125)	(.021)	(.142)	(.145)	(.108)
Modified	.592	.586	.572	.504	NA	NA	NA
	(.121)	(.114)	(.132)	(.131)	NA	NA	NA

NOTE: The error is averaged over five replications The reported estimates have been divided by 10^{49}. The estimates for each replication are based on a Metropolis sampling chain of length 100,000. The numbers in parentheses are the standard errors of the estimates of errors, based on 5 replications.

we take $\hat{C} = h(\hat{\theta})/p(\hat{\theta}|y)$, where

$$p(\hat{\theta}|y) = p(\hat{\theta}_1|y)p(\hat{\theta}_2|y, \hat{\theta}_1) \ldots p(\hat{\theta}_{10}|y, \hat{\theta}_1, \ldots, \hat{\theta}_9).$$

Each term in the foregoing equation was estimated with a separate simulation. Thus we required 10 simulations. Kernel density estimation was then used to estimate each density.

We repeated the analysis five times. In addition, Alan Genz of Washington State University kindly applied four different numerical methods to this integration problem, including subregion-adaptive Monte Carlo and subregion-adaptive quadrature (Genz and Kass 1997), modified Gauss–Hermite (Genz and Keister 1996), and spherical-radial integration rules (Monahan and Genz 1996, 1997). All four methods produced results of $\hat{C} = 2.85 \cdot 10^{49}$ with an apparent error less than .01. Using this estimate as the true value C, Table 3 shows the relative absolute error $|\hat{C} - C|/C$ (averaged over the five runs) for each of our posterior simulation-based estimators. In the table, "bridge (5)" refers to the bridge estimator after five iterations. All of the methods produce reasonable estimates, though postimportance sampling does better than the others. This suggests that the posterior is approximately normal, a hypothesis consistent with results from Genz's analysis.

It is instructive to summarize the amount of work involved in each method. Let p represent the dimension of the parameter space and let m represent the number samples in a simulation (100,000 in the example). Table 4 shows the number of function evaluations and number of simulations required for each method. In our example, the Laplace method requires one simulation (of size $m = 100{,}000$) and a single function evaluation. The bridge sampler requires two simulations (each of size 100,000) and 200,000 function evaluations. The Chib estimator requires 10 simulations (each of size 100,000) and then 10 function evaluations. The Laplace method requires the least effort. Deciding which of the other methods requires the least effort is problem-dependent. If running the simulations is simple, then the Chib method might be the quickest. If the simulations are difficult and require much fine tuning, then bridge sampling might be better. On the other hand, if function evaluations are costly, bridge sampling can be time-consuming. It should be noted that the number of simulations required for the Chib method can sometimes be reduced if the parameters are blocked together appropriately. In the foregoing example, it would make sense to estimate $p(\theta_3, \theta_4, \theta_5, \theta_6|y)$, because the complete conditional for these parameters is Gaussian. This would reduce the number of simulations to seven.

7. CONCLUSION

In this article we have evaluated generic ways of computing Bayes factors via posterior simulation. The methods that we compared are all easily programmed without reference to the particulars of the problem at hand. Thus, for instance, it is possible to write computer programs for these approximations that require only the simulated values of the posterior together with a function that evaluates the product of the likelihood with the prior density. The most important findings from our study are (a) all modified methods are reasonably accurate, and (b) bridge sampling, using the normal approximation to the posterior, often provides substantial improvement; furthermore, Laplace bridge sampling is nearly as accurate as optimal bridge sampling. An important point is that the volume-corrected Laplace approximation $\hat{C}_{\mathrm{L}}^*$ requires only a single evaluation of the posterior density, in contrast to the others, which require thousands of such evaluations. This suggests the use of $\hat{C}_{\mathrm{L}}^*$ whenever evaluation of the posterior density is difficult or expensive.

Our limited study is by no means definitive, but it does suggest specific guidelines. It leads us to recommend the following:

1. If it is likely that the posterior densities in (1) each has a dominant peak in the interior of the parameter space, then the methods in this article should be applicable. The volume-corrected Laplace estimator $\hat{C}_{\mathrm{L}}^*$ should be computed, using $\alpha = .05$ and $m = 100{,}000$; if posterior simulation takes so long that $m = 100{,}000$ is impractical, m should be as large as possible.

   a. If the posterior density is costly, then the relative difference $(\hat{C}_{\mathrm{L}} - \hat{C}_{\mathrm{L}}^*)/\hat{C}_{\mathrm{L}}^*$ should be examined. If it is

Table 4. Comparison of Methods

	Laplace	Bartlett	Reciprocal	Importance	Bridge	Chib
Simulations	1	1	1	2	2	p
Function evaluations	1	m	m	m	$2m$	p

NOTE: Here, p is the dimension of the parameter space, and m is the simulation sample size per run.

small, then $\hat{C}_{\mathrm{L}}^{*}$ is likely to be accurate. Otherwise, other methods should be attempted.

b. If posterior density evaluations are easy to obtain, then the Laplace bridge estimator based on the normal approximation to the posterior should be used. Again, the result may be compared to $\hat{C}_{\mathrm{L}}^{*}$ to roughly assess accuracy, keeping in mind that the bridge estimator may be as much as 10 times more accurate. If accuracy remains dubious, then other methods (see item 2) may be used as a check.

2. If either posterior is likely to be strongly multimodal, or to have a mode on the boundary of the parameter space, then the methods of this are unlikely to be helpful. Alternatives described by Carlin and Chib (1995), Chib (1994), and Green (1995) may be effective, however.

We should add that if very large sample sizes are used for the Monte Carlo, then it may be necessary to let α be a function of m in the modified methods. We also point out that for nested hypotheses, the Savage–Dickey method discussed by Verdinelli and Wasserman (1995) is effective and easily implemented if the parameter being tested is a scalar. Further work is needed to assess this method when the difference in the dimensions of the two models is larger than one.

Markov chain Monte Carlo and other posterior simulation methods often involve the simulation of a large number of parameters, of which many are "latent data" or "random effects" introduced to simplify the algorithm. A direct application of our methods is unlikely to work well with such high-dimensional parameters, as the posterior density will tend to be flat in the direction of the random effects and so not have a dominant mode.

But our methods may still work if attention is restricted to the "fixed effects" or consistently estimable parameters and the simulated values of the random effects are simply discarded. It then remains to evaluate the likelihood, which is now an integrated likelihood for the fixed effects, with the random effects integrated out. In the broad class of conditionally independent hierarchical models (Kass and Steffey 1989), the integrals involved are of low dimension and can be evaluated fairly easily, at least approximately (see Lewis and Raftery 1994 and Raftery 1995).

APPENDIX: THEORETICAL DETAILS OF THE BARTLETT ADJUSTMENT

Let θ be scalar. Let $H(\theta) = \log h(\theta)$, $H_j(\theta) = d^j H(\theta)/d\theta^j$, and $\hat{H}_j = H_j(\hat{\theta})$, $(j = 1, \ldots, p)$; hence $-\hat{H}_2 = \hat{\Sigma}^{-1}$. To error of order $O(n^{-3/2})$,

$$
\begin{aligned}
&C\{h(\hat{\theta})\}^{-1}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad = (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\int_{\Theta} \exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\
&\quad = (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\int_{\hat{\theta}-k}^{\hat{\theta}+k} \exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\
&\quad = (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\int_{-k}^{k} \exp\left\{\frac{1}{2}\, b^2\hat{H}_2 + \frac{1}{6}\, b^3\hat{H}_3 + \frac{1}{24}\, b^4\hat{H}_4\right\} db \\
&\quad = (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\int_{-k}^{k} \exp\left\{-\frac{1}{2}\, b^2\hat{\Sigma}^{-1}\right\} \\
&\qquad \times \exp\left\{\frac{1}{6}\, b^3\hat{H}_3 + \frac{1}{24}\, b^4\hat{H}_4\right\} db \\
&\quad = (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\, b^2\hat{\Sigma}^{-1}\right\} \\
&\qquad \times \left\{1 + \frac{1}{6}\, b^3\hat{H}_3 + \frac{1}{24}\, b^4\hat{H}_4 + \frac{1}{72}\, b^6\hat{H}_3^2\right\} db \\
&\quad = 1 + \frac{1}{8}\,\hat{H}_4\hat{\Sigma}^2 + \frac{5}{24}\,\hat{H}_3^2\hat{\Sigma}^3, \qquad \text{(A.1)}
\end{aligned}
$$

where k is an arbitrary positive constant and $b = \theta - \hat{\theta}$. In carrying out this calculation, essentially three approximations are made: The range of integration is restricted from the entire parameter space Θ to an interval $[\hat{\theta} - k, \hat{\theta} + k]$ of fixed length; the integrand is approximated by a Taylor expansion over the interval; and the domain of integration is enlarged from the interval to the entire real line to facilitate evaluation of the integral. The final integration uses the fact that if $\theta \sim N(\hat{\theta}, \hat{\Sigma})$, then $E\{(\theta - \hat{\theta})^{2j}\} = (2j)!\hat{\Sigma}^j/\{2^j(j)!\}$.

By including further terms in the Taylor approximation, it can be shown that the error in (A.1) is actually of order $O(n^{-2})$. Thus

$$
\begin{aligned}
C &= h(\hat{\theta})(2\pi)^{1/2}\hat{\Sigma}^{1/2}\{1 + \kappa/2 + O(n^{-2})\} \\
&= \hat{C}_{\mathrm{L}}\{1 + \kappa/2 + O(n^{-2})\}, \qquad \text{(A.2)}
\end{aligned}
$$

where

$$
\kappa = \frac{1}{4}\,\hat{H}_4\hat{\Sigma}^2 + \frac{5}{12}\,\hat{H}_3^2\hat{\Sigma}^3 = O(n^{-1}).
$$

Let $W = 2\{H(\hat{\theta}) - H(\theta)\}$. To error of order $O(n^{-3/2})$,

$$
\begin{aligned}
&E(W|y) \\
&\quad = C^{-1}h(\hat{\theta})\int_{\Theta} 2\{H(\hat{\theta}) - H(\theta)\}\exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\
&\quad = C^{-1}h(\hat{\theta})\int_{\hat{\theta}-k}^{\hat{\theta}+k} 2\{H(\hat{\theta}) - H(\theta)\}\exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\
&\quad = C^{-1}h(\hat{\theta})\int_{-k}^{k} \exp\left\{-\frac{1}{2}\, b^2\hat{\Sigma}^{-1}\right\} \\
&\qquad \times \left\{1 + \frac{1}{6}\, b^3\hat{H}_3 + \frac{1}{24}\, b^4\hat{H}_4 + \frac{1}{72}\, b^6\hat{H}_3^2\right\} \\
&\qquad \times \left\{b^2\hat{\Sigma}^{-1} - \frac{1}{3}\, b^3\hat{H}_3 - \frac{1}{12}\, b^4\hat{H}_4\right\} db \\
&\quad = C^{-1}h(\hat{\theta})\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\, b^2\hat{\Sigma}^{-1}\right\} \\
&\qquad \times \left\{b^2\hat{\Sigma}^{-1} - \frac{1}{3}\, b^3\hat{H}_3 + \frac{1}{6}\, b^5\hat{H}_3\hat{\Sigma}^{-1} - \frac{1}{12}\, b^4\hat{H}_4\right. \\
&\qquad \left. - \frac{1}{18}\, b^6\hat{H}_3^2 + \frac{1}{24}\, b^6\hat{H}_4\hat{\Sigma}^{-1} + \frac{1}{72}\, b^8\hat{H}_3^2\hat{\Sigma}^{-1}\right\} db \\
&\quad = C^{-1}h(\hat{\theta})(2\pi)^{1/2}\hat{\Sigma}^{1/2}\left\{1 + \frac{3}{8}\,\hat{H}_4\hat{\Sigma}^2 + \frac{5}{8}\,\hat{H}_3^2\hat{\Sigma}^3\right\} \\
&\quad = 1 + \frac{1}{4}\,\hat{H}_4\hat{\Sigma}^2 + \frac{5}{12}\,\hat{H}_3^2\hat{\Sigma}^3, \qquad \text{(A.3)}
\end{aligned}
$$

because, by (A.1),

$$
C^{-1} = \{h(\hat{\theta})\}^{-1}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\left\{1 - \frac{1}{8}\,\hat{H}_4\hat{\Sigma}^2 - \frac{5}{24}\,\hat{H}_3^2\hat{\Sigma}^3\right\}.
$$

Again, if higher-order terms are taken into account, then it can be shown that the error in (A.3) is order $O(n^{-2})$. Hence

$$E(W|y) = 1 + \kappa + O(n^{-2}), \tag{A.4}$$

and combining (A.2) and (A.4) shows that

$$\begin{aligned} C &= \hat{C}_{\mathrm{L}} \left\{ 1 + \frac{E(W|y) - 1}{2} + O(n^{-2}) \right\} \\ &= \hat{C}_{\mathrm{L}} [\{E(W|y)\}^{1/2} + O(n^{-2})]. \end{aligned} \tag{A.5}$$

Similar calculations apply in the vector case. For $p \geq 1$,

$$C = \hat{C}_{\mathrm{L}} \{1 + \kappa/2 + O(n^{-2})\} \tag{A.6}$$

and

$$E(W|y) = p + \kappa + O(n^{-2}) = p(1 + \kappa/p) + O(n^{-2}), \tag{A.7}$$

with

$$\begin{aligned} \kappa = \frac{1}{4}\, \hat{H}_{abcd} \hat{\Sigma}^{ab} \hat{\Sigma}^{cd} + \frac{1}{4}\, \hat{H}_{abc} \hat{H}_{def} \hat{\Sigma}^{ad} \hat{\Sigma}^{bc} \hat{\Sigma}^{ef} \\ + \frac{1}{6}\, \hat{H}_{abc} \hat{H}_{def} \hat{\Sigma}^{ad} \hat{\Sigma}^{be} \hat{\Sigma}^{cf} = O(n^{-1}), \end{aligned} \tag{A.8}$$

where $H(\theta) = \log h(\theta)$, $H_{ab}(\theta) = \partial^2 H(\theta)/\partial\theta^a \partial\theta^b$, $H_{abc}(\theta) = \partial^3 H(\theta)/\partial\theta^a \partial\theta^b \partial\theta^c$, $\hat{H}_{ab} = H_{ab}(\hat{\theta})$, $\hat{H}_{abc} = H_{abc}(\hat{\theta})$, $(a, b, c = 1, \ldots, p)$, and so on, and $\hat{\Sigma} = (\hat{\Sigma}^{ab}) = (-\hat{H}_{ab})^{-1}$. The general version of (A.5) that emerges from (A.6) and (A.7) is

$$\begin{aligned} C &= \hat{C}_{\mathrm{L}} \left\{ 1 + \frac{E(W|y) - p}{2} + O(n^{-2}) \right\} \\ &= \hat{C}_{\mathrm{L}} \left[\left\{ \frac{E(W|y)}{p} \right\}^{p/2} + O(n^{-2}) \right]. \end{aligned} \tag{A.9}$$

Now let $P_k = P(I_k)$ be the posterior probability content of the interval $I_k = [\hat{\theta} - k, \hat{\theta} + k]$ and let $\Phi_k = \Phi(I_k)$. Taking the interval I_k into account in calculation (A.1) yields

$$\begin{aligned} & P_k C \{h(\hat{\theta})\}^{-1} (2\pi)^{-1/2} \hat{\Sigma}^{-1/2} \\ &= (2\pi)^{-1/2} \hat{\Sigma}^{-1/2} \int_{\hat{\theta}-k}^{\hat{\theta}+k} \exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\ &= (2\pi)^{-1/2} \hat{\Sigma}^{-1/2} \int_{-k}^{k} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} \\ &\quad \times \left\{ 1 + \frac{1}{6}\, b^3 \hat{H}_3 + \frac{1}{24}\, b^4 \hat{H}_4 + \frac{1}{72}\, b^6 \hat{H}_3^2 \right\} db \\ &= (2\pi)^{-1/2} \hat{\Sigma}^{-1/2} \left(\int_{-k}^{k} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} db \right. \\ &\quad \left. + \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} \left\{ \frac{1}{24}\, b^4 \hat{H}_4 + \frac{1}{72}\, b^6 \hat{H}_3^2 \right\} db \right) \\ &= \Phi_k + \frac{1}{8}\, \hat{H}_4 \hat{\Sigma}^2 + \frac{5}{24}\, \hat{H}_3^2 \hat{\Sigma}^3 \\ &= \Phi_k + \kappa/2. \end{aligned} \tag{A.10}$$

The error in (A.10) is of order $O(n^2)$, so that

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{\kappa}{2\Phi_k} + O(n^{-2}) \right\}. \tag{A.11}$$

Furthermore, to error of order $O(n^{-2})$, the conditional posterior expectation of W, given $\theta \in I_k$, is

$$\begin{aligned} & E(W|I_k, y) \\ &= C^{-1} P_k^{-1} h(\hat{\theta}) \int_{\hat{\theta}-k}^{\hat{\theta}+k} 2\{H(\hat{\theta}) - H(\theta)\} \exp\{H(\theta) - H(\hat{\theta})\}\, d\theta \\ &= C^{-1} P_k^{-1} h(\hat{\theta}) \int_{-k}^{k} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} \\ &\quad \times \left\{ 1 + \frac{1}{6}\, b^3 \hat{H}_3 + \frac{1}{24}\, b^4 \hat{H}_4 + \frac{1}{72}\, b^6 \hat{H}_3^2 \right\} \\ &\quad \times \left\{ b^2 \hat{\Sigma}^{-1} - \frac{1}{3}\, b^3 \hat{H}_3 - \frac{1}{12}\, b^4 \hat{H}_4 \right\} db \\ &= C^{-1} P_k^{-1} h(\hat{\theta}) \left(\int_{-k}^{k} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} \{b^2 \hat{\Sigma}^{-1}\}\, db \right. \\ &\quad + \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\, b^2 \hat{\Sigma}^{-1} \right\} \left\{ -\frac{1}{12}\, b^4 \hat{H}_4 - \frac{1}{18}\, b^6 \hat{H}_3^2 \right. \\ &\quad \left. \left. + \frac{1}{24}\, b^6 \hat{H}_4 \hat{\Sigma}^{-1} + \frac{1}{72}\, b^8 \hat{H}_3^2 \hat{\Sigma}^{-1} \right\} db \right) \\ &= C^{-1} P_k^{-1} h(\hat{\theta}) (2\pi)^{1/2} \hat{\Sigma}^{1/2} \left\{ \Phi_k N_k + \frac{3}{8}\, \hat{H}_4 \hat{\Sigma}^2 + \frac{5}{8} \hat{H}_3^2 \hat{\Sigma}^3 \right\} \\ &= C^{-1} \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ N_k + \frac{3\kappa}{2\Phi_k} \right\}, \end{aligned} \tag{A.12}$$

where $N_k = E\{(\theta - \hat{\theta})^2 \hat{\Sigma}^{-1} | \theta \in I_k\}$ with $\theta \sim N(\hat{\theta}, \hat{\Sigma})$.

Now, (A.11) yields

$$C^{-1} = \hat{C}_{\mathrm{L}}^{-1} \frac{P_k}{\Phi_k} \left\{ 1 - \frac{\kappa}{2\Phi_k} + O(n^{-2}) \right\},$$

and substituting this formula into (A.12) produces

$$E(W|I_k, y) = N_k + \frac{\kappa}{2\Phi_k}(3 - N_k) + O(n^{-2}). \tag{A.13}$$

Hence

$$\frac{\kappa}{2\Phi_k} = \frac{E(W|I_k, y) - N_k}{3 - N_k} + O(n^{-2})$$

and, by (A.11),

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{E(W|I_k, y) - N_k}{3 - N_k} + O(n^{-2}) \right\}. \tag{A.14}$$

In the vector case, let $B_k = \{\theta \in \Theta | (\theta - \hat{\theta})' \hat{\Sigma}^{-1} (\theta - \hat{\theta}) \leq nk^2\}$. Under the distribution $\theta \sim N(\hat{\theta}, \hat{\Sigma})$, let $N_k = E\{(\theta - \hat{\theta})' \hat{\Sigma}^{-1} (\theta - \hat{\theta}) | \theta \in B_k\}$. By arguments similar to those used for calculating the integrals in (A.10) and (A.12), the general versions of (A.11) and (A.13) are found to be

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{\kappa}{2\Phi_k} + O(n^{-2}) \right\}$$

and

$$E(W|I_k, y) = N_k + \frac{\kappa}{2\Phi_k} \{p + 2 - N_k\} + O(n^{-2}),$$

where κ is given by (A.8). Hence

$$C = \hat{C}_{\mathrm{L}} \frac{\Phi_k}{P_k} \left\{ 1 + \frac{E(W|I_k, y) - N_k}{p + 2 - N_k} + O(n^{-2}) \right\}.$$

This expression for C motivates (5), the local version of the Bartlett-adjusted Laplace estimator. Note that $N_k = p(\Pr\{\chi^2_{p+2} \leq nk^2\} / \Pr\{\chi^2_p \leq nk^2\})$.

In the case where δ is fixed, we get

$$
\begin{aligned}
&P_k C\{h(\theta)\}^{-1}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&= (2\pi)^{-1/2}\hat{\Sigma}^{-1/2}\left(\int_{-k}^{k}\exp\left\{-\frac{1}{2}\,b^2\hat{\Sigma}^{-1}\right\}db\right. \\
&\quad\left.+\int_{-k}^{k}\exp\left\{-\frac{1}{2}\,b^2\hat{\Sigma}^{-1}\right\}\left\{\frac{1}{24}\,b^4\hat{H}_4+\frac{1}{72}\,b^6\hat{H}_3^2\right\}db\right) \\
&= \Phi_k\left\{1+\frac{1}{8}\,\hat{H}_4\hat{\Sigma}^2+\frac{5}{24}\,\hat{H}_3^2\hat{\Sigma}^3\right\} \\
&\quad-\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left\{\hat{H}_4\left(\frac{1}{12}\,k^3\hat{\Sigma}+\frac{1}{4}\,k\hat{\Sigma}^2\right)\right. \\
&\quad\left.+\hat{H}_3^2\left(\frac{1}{36}\,k^5\hat{\Sigma}+\frac{5}{36}\,k^3\hat{\Sigma}^2+\frac{5}{12}\,k\hat{\Sigma}^3\right)\right\}, \qquad \text{(A.15)}
\end{aligned}
$$

with error of order $O(n^{-2})$. Rearrangement of (A.15) gives

$$
C=\hat{C}_{\mathrm{L}}\,\frac{\Phi_k}{P_k}\left\{1+\frac{\kappa'}{2}+O(n^{-2})\right\}, \qquad \text{(A.16)}
$$

where

$$
\begin{aligned}
\kappa' &= \frac{1}{4}\,\hat{H}_4\hat{\Sigma}^2+\frac{5}{12}\,\hat{H}_3^2\hat{\Sigma}^3-\Phi_k^{-1} \\
&\quad\times\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left\{\hat{H}_4\left(\frac{1}{6}\,k^3\hat{\Sigma}+\frac{1}{2}\,k\hat{\Sigma}^2\right)\right. \\
&\quad\left.+\hat{H}_3^2\left(\frac{1}{18}\,k^5\hat{\Sigma}+\frac{5}{18}\,k^3\hat{\Sigma}^2+\frac{5}{6}\,k\hat{\Sigma}^3\right)\right\}.
\end{aligned}
$$

Similarly, the appropriate version of (A.12) is

$$
\begin{aligned}
&E(W|I_k,y) \\
&= C^{-1}P_k^{-1}h(\hat{\theta})\left(\int_{-k}^{k}\exp\left\{-\frac{1}{2}\,b^2\hat{\Sigma}^{-1}\right\}\{b^2\hat{\Sigma}^{-1}\}\,db\right. \\
&\quad\left.+\int_{-k}^{k}\exp\left\{-\frac{1}{2}\,b^2\hat{\Sigma}^{-1}\right\}\left\{-\frac{1}{12}\,b^4\hat{H}_4-\frac{1}{18}\,b^6\hat{H}_3^2\right.\right. \\
&\quad\left.\left.+\frac{1}{24}\,b^6\hat{H}_4\hat{\Sigma}^{-1}+\frac{1}{72}\,b^8\hat{H}_3^2\hat{\Sigma}^{-1}\right\}db\right) \\
&= C^{-1}\hat{C}_{\mathrm{L}}\,\frac{\Phi_k}{P_k}\left[N_k+\frac{3}{8}\,\hat{H}_4\hat{\Sigma}^2+\frac{15}{24}\,\hat{H}_3^2\hat{\Sigma}^3-\Phi_k^{-1}\right. \\
&\quad\times\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left\{\hat{H}_4\left(\frac{1}{12}\,k^5+\frac{1}{4}\,k^3\hat{\Sigma}+\frac{3}{4}\,k\hat{\Sigma}^2\right)\right. \\
&\quad\left.\left.+\hat{H}_3^2\left(\frac{1}{36}\,k^7+\frac{1}{12}\,k^5\hat{\Sigma}+\frac{5}{12}\,k^3\hat{\Sigma}^2+\frac{5}{4}\,k\hat{\Sigma}^3\right)\right\}\right] \\
&= C^{-1}\hat{C}_{\mathrm{L}}\,\frac{\Phi_k}{P_k}\left\{N_k+\frac{3\kappa'}{2}-\Phi_k^{-1}\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\right. \\
&\quad\left.\times\hat{\Sigma}^{-1/2}\left(\frac{1}{12}\,\hat{H}_4k^5+\frac{1}{36}\,\hat{H}_3^2k^7\right)\right\} \qquad \text{(A.17)}
\end{aligned}
$$

with error of order $O(n^{-2})$. Note that in this case

$$
N_k=1-2k\Phi_k^{-1}\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2}.
$$

It follows from (A.16) and (A.17) that

$$
\begin{aligned}
E(W|I_k,y) &= N_k+\frac{\kappa'}{2}\,(3-N_k)-\Phi_k^{-1} \\
&\quad\times\exp\{-k^2\hat{\Sigma}^{-1}/2\}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left(\frac{1}{12}\,\hat{H}_4k^5+\frac{1}{36}\,\hat{H}_3^2k^7\right)+O(n^{-2}) \\
&= N_k+\frac{\kappa'}{2}\,(3-N_k)-k\Phi_k^{-1}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left[\frac{h(\hat{\theta}+k)+h(\hat{\theta}-k)}{h(\hat{\theta})}-2\exp\{-k^2\hat{\Sigma}^{-1}/2\}\right] \\
&\quad+O(n^{-2}).
\end{aligned}
$$

Hence

$$
\frac{\kappa'}{2}=\frac{E(W|I_k,y)+D_k-N_k}{3-N_k}+O(n^{-2}),
$$

where

$$
\begin{aligned}
D_k &= k\Phi_k^{-1}(2\pi)^{-1/2}\hat{\Sigma}^{-1/2} \\
&\quad\times\left[\frac{h(\hat{\theta}+k)+h(\hat{\theta}-k)}{h(\hat{\theta})}-2\exp\{-k^2\hat{\Sigma}^{-1}/2\}\right] \qquad \text{(A.18)}
\end{aligned}
$$

and, by (A.14),

$$
C=\hat{C}_{\mathrm{L}}\,\frac{\Phi_k}{P_k}\left\{1+\frac{E(W|I_k,y)+D_k-N_k}{3-N_k}+O(n^{-2})\right\}. \qquad \text{(A.19)}
$$

[Received July 1995. Revised October 1996.]

REFERENCES

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Bates, D., and Watts, D. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: Wiley.

Bennett, C. (1976), "Efficient Estimation of Free Energy Differences From Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science*, 10, 3–66.

Carlin, B., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society*, Ser. B, 57, 473–484.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

DiCiccio, T. J., and Stern, S. E. (1993), "On Bartlett Adjustments for Approximate Bayesian Inference," *Biometrika*, 80, 731–740.

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501–514.

Gelman, A., and Meng, X. (1994), "Path Sampling for Computing Normalizing Constants: Identities and Theory," unpublished manuscript.

Genz, A. and Kass, R. E. (in press), "Subregion-Adaptive Integration of Integrals having a dominant peak," *Journal of Computational and Graphical Statistics*, 6.

Genz, A., and Keister, B., (1996), "Fully Symmetric Interpolatory Rules for Multiple Integrals over Infinite Regions With Gaussian Weight," *Journal of Computational Applied Mathematics*, 71, 299–309.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1340.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Gnanadesikan, R. (1976), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: Wiley.

Green, P. (1995), "Reversible Jump MCMC Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hammersley, J. M., and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Chapman and Hall.

Hjort, N., and Glad, I. (1995), "Nonparametric Density Estimation With a Parametric Start," *The Annals of Statistics*, 23, 882–904.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 377–395.

Kass, R. E., and Steffey, D. L. (1989), "Approximate Bayesian Methods in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," *Journal of the American Statistical Association*, 84, 717–726.

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

Lewis, S., and Raftery, A. E. (1994), "Estimating Bayes Factors Via Posterior Simulation with the Laplace–Metropolis Estimator," Technical report 279, University of Washington, Dept. of Statistics.

Monahan, J., and Genz, A. (in press), "A Comparison of Omnibus Methods for Bayesian Computation," *Computing Science and Statistics*, 27.

Monahan, J., and Genz, A. (1997), "Spherical-Radial Integration Rules for Bayesian Computation," *Journal of the American Statistical Association*, 92, 664–674.

Meng, X. L., and Wong, W. H. (1993), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," Technical Report 365, University of Chicago, Dept. of Statistics.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 56, 3–48.

O'Hagan, A., and Leonard, T. (1976), "Bayes Estimation Subject to Uncertainty About Parameter Constraints," *Biometrika*, 63, 201–203.

Phillips, D. B., and Smith, A. F. M. (1994), "Bayesian Model Comparison via Jump Diffusion," Technical Report 94-20, Imperial College London, Dept. of Mathematics.

Raftery, A. E. (1995), "Hypothesis Testing and Model Selection via Posterior Simulation," in *Practical Markov Chain Monte Carlo*, eds. W. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 163–188.

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points" (with discussion), *Journal of the American Statistical Association*, 85, 633–651.

Rubin, D. B. (1987), "Comment on 'The Calculation of Posterior Distributions by Data Augmentation,' " by M. A. Tanner and W. H. Wong, *Journal of the American Statistical Association*, 82, 543–546.

——— (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 395–402.

Smith, A. F. M., and Roberts, G. (1992), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Methods" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors Using a Generalization of the Savage–Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.