Estimating Bayes Factors via Posterior Simulation with the Laplace—Metropolis Estimator

Steven M. Lewis & Adrian E. Raftery

Estimating Bayes Factors via Posterior Simulation With the Laplace–Metropolis Estimator

Steven M. LEWIS and Adrian E. RAFTERY

The key quantity needed for Bayesian hypothesis testing and model selection is the integrated, or marginal, likelihood of a model. We describe a way to use posterior simulation output to estimate integrated likelihoods. We describe the basic Laplace—Metropolis estimator for models without random effects. For models with random effects, we introduce the compound Laplace–Metropolis estimator. We apply this estimator to data from the World Fertility Survey and show it to give accurate results. Batching of simulation output is used to assess the uncertainty involved in using the compound Laplace–Metropolis estimator. The method allows us to test for the effects of independent variables in a random-effects model and also to test for the presence of the random effects.

KEY WORDS: Compound Laplace–Metropolis estimator; Integrated likelihood; Marginal likelihood; Markov chain Monte Carlo; Random-effects model; World Fertility Survey.

1. INTRODUCTION

The standard Bayesian solution to the hypothesis testing and model selection problems is to compute Bayes factors for comparing two or more competing models (see Kass and Raftery 1995 for a survey). The Bayes factor, B_{01}, for comparing model M_0 to model M_1 for observed data, $\mathbf{Y}$, is the ratio of the posterior odds for M_0 against M_1 to the prior odds, which reduces to

$$B_{01} = \frac{f(\mathbf{Y}|M_0)}{f(\mathbf{Y}|M_1)}.$$

In other words, the Bayes factor is the ratio of the integrated (or marginal) likelihoods of the two models being compared. Hence calculation of Bayes factors boils down to computing integrated likelihoods,

$$f(\mathbf{Y}|M_m) = \int f(\mathbf{Y}|\boldsymbol{\theta}_m, M_m) f(\boldsymbol{\theta}_m|M_m) \, d\boldsymbol{\theta}_m \qquad m = 0, 1,$$

where $\boldsymbol{\theta}_m$ is the vector of parameters in model M_m and $f(\boldsymbol{\theta}_m|M_m)$ is its prior density. Dropping the notational dependence on the model, this can be rewritten as

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \qquad (1)$$

Historically, the integration required for calculating integrated likelihoods has been done by taking advantage of conjugacy, by assuming approximate posterior normality, or by using numerical quadrature, the Laplace method, or Monte Carlo integration (see Kass and Raftery 1995 for a review and references). More recently, it has become possible to estimate a wider range of models than previously, using posterior simulation methods such as Markov chain Monte Carlo (MCMC), the sampling importance resampling (SIR) algorithm of Rubin (1987, 1988), and the weighted likelihood bootstrap (Newton and Raftery 1994). Previous ways of calculating integrated likelihoods often cannot be used for models estimated via MCMC or other posterior simulation methods.

Several ways of estimating integrated likelihoods from posterior simulation output have been suggested. These were surveyed by Raftery (1996a) and include importance sampling methods such as the harmonic mean of the output likelihoods and modifications thereof (Newton and Raftery 1994), bridge sampling (Meng and Wong 1993), and path sampling (Gelman and Meng 1994) (see also Gelfand and Dey 1994). For the case where all complete conditional densities have closed-form expressions, Chib (1995) has described a method for finding integrated likelihoods from posterior simulation output. In many applications these closed-form densities are not available, in which case this method will not be applicable.

Carlin and Chib (1993), George and McCulloch (1993), Green (1995), and Madigan and York (1995) have proposed MCMC methods that move through the set of models considered, eventually visiting each one with a frequency proportional to its posterior model probability. These methods can be used to estimate Bayes factors, but they require that a new MCMC method be designed either in addition to or incorporating the one used to estimate the models. Our focus here is on the computation of integrated likelihoods, and hence Bayes factors, from the posterior simulation output for the individual models.

In Section 3 we describe the basic Laplace–Metropolis estimator of the integrated likelihood. In hierarchical models with both fixed and random effects, the basic Laplace–Metropolis estimator cannot be used directly. Thus, in Section 4 we introduce the compound Laplace–Metropolis estimator for hierarchical models. This estimator results from applying the Laplace method at two different levels; it is applied at a second level to integrate out each of the random-effects parameters.

In Section 6 we use the compound Laplace–Metropolis estimator to calculate log-integrated likelihoods for a number of different models fit to data collected in Iran as part of the World Fertility Survey. By taking the difference be-

tween log-integrated likelihoods for two competing models, we can estimate the log Bayes factor for comparison of the two models.

How variable are the log-integrated likelihood estimates produced by the compound Laplace–Metropolis estimator? In Section 7 we show how to answer this question by batching the posterior simulation output. We close by comparing the estimates of the log-integrated likelihoods provided by the compound Laplace–Metropolis estimator with a "gold standard" based on very large samples from the prior distribution. The compound Laplace–Metropolis estimates are seen to be remarkably accurate, particularly in light of the substantially reduced computing time required.

2. CALCULATING THE INTEGRATED LIKELIHOOD USING THE PRIOR DISTRIBUTION

Before describing the Laplace–Metropolis estimator, we first describe a method for obtaining a "gold standard" for the accurate calculation of integrated likelihoods with which we can compare the Laplace–Metropolis estimator developed in this article.

Because the integral in Equation (1) is not generally analytically tractable, it is usually necessary to approximate it. This integral can be approximated by a simple Monte Carlo estimator of the form

$$\hat{f}_{mc} = \frac{1}{J} \sum_{j=1}^{J} f(\mathbf{Y}|\boldsymbol{\theta}^{(j)}), \qquad (2)$$

where $\{\boldsymbol{\theta}^{(j)}: j = 1, \ldots, J\}$ is a sample from the *prior* distribution of the parameters. Unfortunately, it is generally necessary to obtain a very large number of draws from the prior distribution before $\hat{f}_{mc}$ becomes a good estimator for the integral in Equation (1). In one study, Lewis (1994) found that to reduce the Monte Carlo standard error to an acceptable level, it was necessary to use a sample of roughly 50 million draws from the prior distribution. The computing effort needed to obtain such a large sample can be enormous, requiring several days of computer time on most workstations currently available. In this article we describe an estimator of the integrated likelihood that needs much less computer time.

However, to check the accuracy of the estimator that we describe, we have also calculated the Monte Carlo estimate using Equation (2) for some of the models demonstrated in this article; we refer to this Monte Carlo estimate as the "correct" or "actual" log integrated likelihood. Unlike the other methods considered, this has the advantage of being the mean of a large number of independent and identically distributed random variables, to which the law of large numbers and the central limit theorem apply. Thus, given enough samples from the prior, we can be sure of getting close to the correct value of the integrated likelihood and also of having a reliable statement of the error involved.

3. THE LAPLACE-METROPOLIS ESTIMATOR OF THE MARGINAL LIKELIHOOD

The Laplace approximation for an integral of the form

$\int e^{h(\mathbf{u})} d\mathbf{u}$ is found using a Taylor series expansion of a real-valued function $h(\mathbf{u})$ of a P-dimensional vector $\mathbf{u}$. The resulting Laplace approximation is

$$\int e^{h(\mathbf{u})} d\mathbf{u} \approx (2\pi)^{P/2} |\mathbf{H}^*|^{1/2} \exp\{h(\mathbf{u}^*)\},$$

where $\mathbf{u}^*$ is the value of $\mathbf{u}$ at which h attains its maximum and $\mathbf{H}^*$ is minus the inverse Hessian of h evaluated at $\mathbf{u}^*$. The Laplace approximation is justified when h is a smooth, bounded unimodal function, with a single dominant mode at $\mathbf{u}^*$ (Tierney and Kadane 1986). In practice the Laplace approximation often works well even for some functions that do not completely satisfy these conditions. We present an example of this later.

For our purposes, we want to use the Laplace method to approximate the integrated likelihood

$$f(\mathbf{Y}) = \int f(\boldsymbol{\theta}) f(\mathbf{Y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta},$$

where $f(\boldsymbol{\theta})$ is the prior distribution of a vector parameter $\boldsymbol{\theta}$ and $f(\mathbf{Y}|\boldsymbol{\theta})$ is the likelihood. Letting $h(\boldsymbol{\theta}) \equiv \log\{f(\boldsymbol{\theta}) f(\mathbf{Y}|\boldsymbol{\theta})\}$, we can apply the Laplace method to derive the following approximation for the integrated likelihood:

$$f(\mathbf{Y}) \approx (2\pi)^{P/2} |\mathbf{H}^*|^{1/2} f(\boldsymbol{\theta}^*) f(\mathbf{Y}|\boldsymbol{\theta}^*), \qquad (3)$$

where $\boldsymbol{\theta}^*$ is the value of $\boldsymbol{\theta}$ at which h attains its maximum and $\mathbf{H}^*$ is minus the inverse Hessian of h evaluated at $\boldsymbol{\theta}^*$. For numerical reasons, and because it is customary to work with log-likelihoods, it is better to work with this approximation on the logarithmic scale. Taking logarithms, we can rewrite Equation (3) as

$$\log\{f(\mathbf{Y})\} \approx \frac{P}{2} \log\{2\pi\} + \frac{1}{2} \log\{|\mathbf{H}^*|\} \\ + \log\{f(\boldsymbol{\theta}^*)\} + \log\{f(\mathbf{Y}|\boldsymbol{\theta}^*)\}. \qquad (4)$$

If $\boldsymbol{\theta}^*$ and $\mathbf{H}^*$ can be found analytically, then we can use Equation (4) to estimate the log-integrated likelihood directly. In many practical situations an analytic solution is not available. Raftery (1996a) has suggested a way to use the Metropolis–Hastings algorithm (Hastings 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) to find estimates for $\boldsymbol{\theta}^*$ and $\mathbf{H}^*$, which may in turn be used in equation (4). We refer to this as the Laplace–Metropolis estimator.

To calculate the Laplace–Metropolis estimator requires Metropolis estimates for both $\boldsymbol{\theta}^*$ and $\mathbf{H}^*$. Assume that we have a posterior sample of parameter simulations from a Metropolis run. There are several possible ways of estimating $\boldsymbol{\theta}^*$ from the sample, including the following:

- Estimate $\boldsymbol{\theta}^*$ as that $\boldsymbol{\theta}$ in the sample at which $h(\boldsymbol{\theta})$ achieves its maximum.
- Estimate the components of $\boldsymbol{\theta}^*$ by finding the componentwise posterior means.
- Estimate the components of $\boldsymbol{\theta}^*$ by finding the componentwise posterior medians.
- Estimate $\boldsymbol{\theta}^*$ by finding the multivariate median, or L_1 center, which is defined as that value of $\boldsymbol{\theta}^{(j)}$ that min-

imizes

$$d(\boldsymbol{\theta}^{(j)}) \equiv \sum_{l=1}^{J} |\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^{(j)}|,$$

where $|\cdot|$ denotes L_1 distance.

- If the dimension of the parameter space is not too large, then the posterior mode might be directly estimated from the posterior sample using nonparametric density estimation.

The first of these methods is the simplest conceptually and usually the most accurate. However, it involves calculating the likelihood many times and so may take a prohibitive amount of computer time. When this happens, we use the multivariate median instead, because it does not require too much computer time and because it provides robustness against the outliers and distance excursions to which MCMC trajectories are prone. We prefer it to the estimated posterior mean because the latter is not robust to outliers, and because in one dimension the median is closer to the mode than is the mean for a wide range of distributions (Johnson and Kotz 1985, pp. 365–366).

The other quantity needed for the Laplace–Metropolis estimator is $\mathbf{H}^*$. This is asymptotically equal to the posterior variance matrix, and we could estimate it by the sample covariance matrix of the posterior simulation output. However, because MCMC trajectories take occasional distant excursions, it is better to use a robust estimator of the posterior variance matrix. For $\mathbf{H}^*$, we use the weighted variance matrix estimate with weights based on the minimum volume ellipsoid estimate of Rousseeuw and van Zomeren (1990).

4. THE COMPOUND LAPLACE–METROPOLIS ESTIMATOR FOR HIERARCHICAL MODELS

We now consider the estimation of integrated likelihoods for comparing alternative hierarchical (i.e., random-effects) models. Here we use the term hierarchical model to mean a model in which several observations are made on subjects, which themselves are grouped into higher-level entities to take account of the aggregation in the sampling design (Raudenbush 1988). Hierarchical models usually involve many nuisance parameters—the random effects—and as a result, calculating integrated likelihoods can be hard. In this situation we have obtained good results by adapting the Laplace–Metropolis estimator to hierarchical models.

For a random-effects model we separate the vector of all parameters into its components: the vector of fixed effects, which we denote by η; the variance of the random-effects hyperparameter, which we denote by Σ; and the vector of random-effects parameters, α. For calculating the Laplace–Metropolis estimator, we need to distinguish between the nuisance parameters, α, and the rest of the parameters and hyperparameters, which together we denote by $\boldsymbol{\theta} \equiv (\eta, \Sigma)$.

The term in the Laplace–Metropolis estimator, Equation (4), requiring additional attention for a random-effects model is the log-likelihood, $\log\{f(\mathbf{Y}|\boldsymbol{\theta}^*)\}$. The first three terms in Equation (4) can be calculated as in Section 3. We still need to find the posterior mode of $\boldsymbol{\theta}$. To do so, we lo-

cate the L_1 center of (η, Σ). The posterior variance matrix of the fixed effects, η, (Σ is not needed) can still be used for $\mathbf{H}^*$. The third term is the logarithm of the joint prior density of the fixed-effects parameters and the variance hyperparameter.

In many random-effects models the random effects are assumed to be conditionally independent given the other parameters in the model, such as the fixed-effects parameters and the hyperparameters. We can take advantage of this assumption to calculate the log-likelihood as simply the sum of the log-likelihoods for each of the random effects. In other words,

$$\log\{f(\mathbf{Y}|\eta, \Sigma)\} = \sum_{i=1}^{I} \log\{f(\mathbf{Y}_i|\eta, \Sigma)\}, \qquad (5)$$

where

$$f(\mathbf{Y}_i|\eta, \Sigma) = \int f(\mathbf{Y}_i|\alpha_i, \eta, \Sigma) f(\alpha_i|\eta, \Sigma) \, d\alpha_i, \qquad (6)$$

with α_i being the random-effects parameter for the ith context.

Equation (5) holds for any (η, Σ). In particular, it holds for $\boldsymbol{\theta}^*$, the joint mode of the posterior distribution, $f(\mathbf{Y}_i|\alpha_i, \eta, \Sigma)$. As a result, we can calculate each of the log-likelihood terms in Equation (5) by conditioning on $\boldsymbol{\theta}^*$.

It remains to calculate the integrals on the right side of Equation (6), each of which will be of low dimension. For regular statistical models, these integrals can be well approximated using the Laplace method as long as there are enough observations on each random effect. Although the Laplace approximation is exact asymptotically only as the number of observations for each random effect becomes large, we show in Section 8 that for the model considered it is very accurate even when there is only one observation per random effect. By using the Laplace method for each of these integrals individually, we obtain the Laplace–Metropolis estimator of the log-integrated likelihood for a general hierarchical model. If we denote the Laplace estimate of the log-conditional likelihood for the ith random effect, as defined in Equation (6), by $\hat{\mathcal{L}}_i$, so that

$$\hat{\mathcal{L}}_i = \log\left\{\int f(\mathbf{Y}_i|\alpha_i, \tilde{\eta}, \tilde{\Sigma}) f(\alpha_i|\tilde{\eta}, \tilde{\Sigma}) \, d\alpha_i\right\},$$

then the Laplace–Metropolis estimator of the log-integrated likelihood for a general hierarchical model, which we refer to as a compound Laplace–Metropolis estimator, is

$$\widehat{\mathcal{LM}}_C = \frac{P}{2} \log\{2\pi\} + \frac{1}{2} \log\{|\mathbf{H}^*|\}$$
$$+ \log\{f(\boldsymbol{\theta}^*)\} + \sum_{i=1}^{I} \hat{\mathcal{L}}_i. \quad (7)$$

5. THE LAPLACE–METROPOLIS ESTIMATOR FOR THE LOGISTIC HIERARCHICAL MODEL

In this section we derive the Laplace–Metropolis estimator for one of the most frequently used types of hierarchical model: the logistic hierarchical model. Here we assume that

the data are produced by a mixed logistic model. That is, we assume that

$$\text{logit}(\pi) = \mathbf{X}\boldsymbol{\eta} + \boldsymbol{\alpha}, \tag{8}$$

where the data may take on only values 0 or 1, $\pi = \{\pi_{it}\}$, π_{it} is the probability that y_{it}, the tth observation within the ith random effect, is a 1, and $\mathbf{X}$ is a matrix of covariate information. The likelihood of the vector of observations for the ith random effect, $\mathbf{Y}_i$, is

$$f(\mathbf{Y}_i|\boldsymbol{\eta}, \alpha_i) = \prod_t \frac{\exp\{\mathbf{X}_{it}\boldsymbol{\eta} + \alpha_i\}^{y_{it}}}{1 + \exp\{\mathbf{X}_{it}\boldsymbol{\eta} + \alpha_i\}}. \tag{9}$$

We take the prior distribution of each of the random-effects parameters to be Gaussian with mean 0 and variance Σ and independent of the fixed-effects parameters and the other random-effects parameters. Then the conditional likelihood of the ith random effect's vector of observations is

$$f(\mathbf{Y}_i|\tilde{\boldsymbol{\eta}}, \tilde{\Sigma}) = \left(\frac{1}{2\pi\tilde{\Sigma}}\right)^{1/2} \int_{-\infty}^{\infty} \exp\{h_2(\alpha_i)\}\, d\alpha_i, \tag{10}$$

where

$$h_2(\alpha_i) = \left\{ \left[\sum_t y_{it}(\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i)\right] - \left(\frac{1}{2\tilde{\Sigma}}\right)\alpha_i^2 \right.$$
$$\left. - \left[\sum_t \log(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i\})\right] \right\}.$$

The first and second derivatives of $h_2(\alpha_i)$ are

$$h_2'(\alpha_i) = \left\{ \left[\sum_t y_{it}\right] - \left(\frac{1}{\tilde{\Sigma}}\right)\alpha_i \right.$$
$$\left. - \left[\sum_i \frac{\exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i\}}{(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i\})}\right] \right\}$$

and

$$h_2''(\alpha_i) = -\left\{ \left(\frac{1}{\tilde{\Sigma}}\right) + \left[\sum_t \frac{\exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i\}}{(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha_i\})^2}\right] \right\}.$$

We can then locate the mode of $h_2(\alpha_i)$, which we denote by α^*, using a few iterations of Newton's method. Using the second derivative above, we find that the square root of the determinant of minus the inverse of the Hessian conditional on θ^* evaluated at α^* is

$$|\mathbf{H}_2|^{1/2} = \tilde{\Sigma}^{1/2}\left(1 + \tilde{\Sigma}\left[\sum_t \frac{\exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*\}}{(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*\})^2}\right]\right)^{-1/2}.$$

We now have all of the quantities needed to use the Laplace method to approximate the integrals on the right side of Equation (6). Doing so, we find that a Laplace estimate for each random-effects log-conditional likelihood is

$$\hat{\mathcal{L}}_i = -\frac{1}{2}\log\left(1 + \tilde{\Sigma}\left[\sum_t \frac{\exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*\}}{(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*\})^2}\right]\right)$$
$$- \left(\frac{1}{2\tilde{\Sigma}}\right)\alpha^{*2} + \left[\sum_t y_{it}(\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*)\right]$$
$$- \left[\sum_t \log(1 + \exp\{\mathbf{X}_{it}\tilde{\boldsymbol{\eta}} + \alpha^*\})\right].$$

These Laplace estimates of the log-conditional likelihoods may then be used in Equation (7) to calculate a compound Laplace–Metropolis estimate for hierarchical models. In the next section we show how this works for an example using data collected in Iran as part of the World Fertility Survey.

6. EXAMPLE USING DATA FROM THE WORLD FERTILITY SURVEY

The Iran Fertility Survey (IFS) was a part of the World Fertility Survey (WFS). The volume edited by Cleland and Scott (1987) serves as the primary summary publication on the WFS. The IFS included full fertility histories of a randomly selected sample of 4,928 married women born between 1926 and 1963. The survey also obtained a large collection of covariate information for each woman, including data on years of formal education for the women as well as their husbands. (For analyses of the full data set, see Raftery, Lewis and Aghajanian 1995 and Raftery, Lewis, Aghajanian, and Kahn, 1996.)

We investigated the methods described in the preceding section by fitting a small example logistic hierarchical model to a randomly selected sample of 29 women from the IFS dataset. The response used in this analysis was whether or not a woman experienced a birth in each year in which she potentially could have had a child; we refer to these as exposure-years.

The first model consisted of four fixed-effect parameters in addition to the intercept. The four fixed effects were the age of the woman during each exposure-year (centered at the average age in the IFS dataset), an indicator variable which is 1 for the first exposure-year of the interval and 0 otherwise, the woman's parity (i.e., number of previous children born) during each exposure-year and the woman's completed education level (a six-level categorical variable). In this example we also included a random-effects parameter for each woman in the sample.

We implemented a Metropolis algorithm for estimating the parameters of a mixed logistic model, Equation (8), in a Fortran program written specifically to handle event history data (Lewis 1993, 1994; Raftery et al. 1995). We obtained the results shown in Table 1. The Metropolis algorithm was run for a total of 5,500 iterations, of which the first 500 were discarded for "burn-in."

Using the technique described in the previous section, we found the estimated log integrated likelihood for this example to be -220.5. How good is this estimate? The actual log-integrated likelihood is -221.8, calculated as in

*Table 1. Bayesian Estimates via MCMC for the Hierarchical
Event History Model for a Sample From the IFS Dataset*

Variable	Posterior mean	Posterior s.d.
Intercept	−.73	.50
Centered Age	−.19	.36
First Year of Interval	−2.34	.45
Parity	−.04	.10
Woman's Education Level	−.35	.20
Variance of Random Effects	.17	.22

Section 2. Thus in this case the compound Laplace–Metropolis estimate is close enough to the true value so as not to be misleading on Jeffreys's (1961, app. B) qualitative scale for the interpretation of Bayes factors (see also Kass and Raftery 1995). In the next section we describe a way of assessing the uncertainty in the estimated integrated likelihood that does not require the very time-consuming calculation of the correct value.

7. ASSESSING THE VARIANCE OF THE ESTIMATOR USING BATCHING

To assess the uncertainty of the Laplace–Metropolis estimator, we use the method of batch means (Geyer 1992; Hastings 1970; Schmeiser 1982). Here we show how it can be applied to the logistic hierarchical model. The idea is to divide an entire MCMC sample into a fairly small number of batches, say B, of equal size. The mean of the simulations within each batch is found, producing B estimates of the mean. Under relatively mild conditions, these B estimates will be essentially independent. The sample of the B estimates can then be used to provide an overall estimate of the mean along with an estimate of its variance.

We calculate separate compound Laplace–Metropolis estimates within each of B batches, take the mean of the B estimates as our overall compound Laplace–Metropolis estimate, and take the variance of the mean as an estimate of the variance of the compound Laplace–Metropolis estimate. In other words, within the bth batch we find the compound Laplace–Metropolis estimate of the log-integrated likelihood, $\widehat{\mathcal{LM}}_b$, as in Equation (7), and then use these B estimates to calculate the overall compound Laplace–Metropolis estimate,

$$\widehat{\mathcal{LM}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathcal{LM}}_b.$$

We applied this method to the model estimated in Section 6. Instead of only 5,500 iterations, we ran the Metropolis algorithm for 75,500 iterations. Once again, we discarded the first 500 as "burn-in." We divided the remaining 75,000 iterations into $B = 15$ batches of 5,000 iterations each.

Before examining parameter estimates or performing other inference using Metropolis, it is a good idea to look at plots of the (dependent) sequential realizations of all the fixed-effects parameter estimates and plots of at least some of the random-effects parameter realizations. We have found that if the Markov chain is not mixing well or is not sampling from the stationary distribution, this is usually apparent in sequential plots of one or more of the fixed-effects
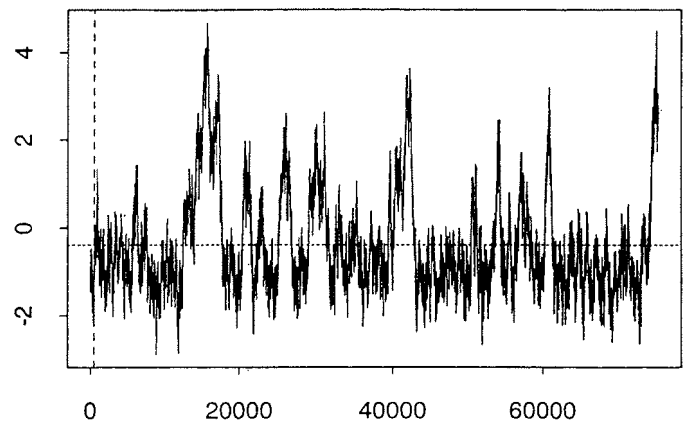

Figure 1. Sequential Realizations of the Intercept Parameter.

realizations. The sequential plot of the intercept realizations is the plot that most often exhibits difficulties in the Markov chain. Figure 1 shows the sequential realizations of the intercept parameter for the four-fixed-effects model. The sequential plots of the other fixed effects were similar to the intercept plot. In this case the Markov chain seems to be mixing well enough and is likely to be sampling from the stationary distribution.

Figure 2 shows the marginal distribution of the intercept parameter; this was obtained using Terrell's (1990) maximal smoothing nonparametric density estimation procedure. This marginal distribution is not normally distributed but is skewed to the right. Thus the normal approximation to this posterior marginal distribution would be quite misleading.

Table 2 shows the estimated log-integrated likelihood within each batch. The first batch consisted of the 501st through 5,500th iterations. The second batch was made up of iterations 5,501–10,500, and so forth. Shown are the contribution of the within-batch maximized log-likelihood, the contribution of the other three terms of Equation (4) to each within-batch log-integrated likelihood, and in the rightmost column the within-batch log-integrated likelihood; the predominant contribution of the maximized log-likelihood is apparent.

The mean of the 15 within-batch estimates is −220.5, and their standard deviation is .7. It can be argued (see Lewis 1994) that the compound Laplace–Metropolis estimator will
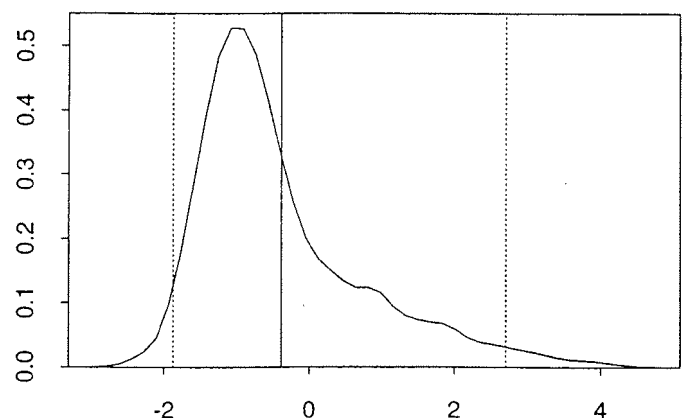

Figure 2. Estimated Marginal Distribution of the Intercept Parameter.

Table 2. Compound Laplace–Metropolis Estimates for Seperate Batches Using the Posterior Simulation Output

Batch	Within-batch log-likelihood $\sum_{i=1}^{I} \hat{\mathcal{L}}_i$	Within-batch other terms	Within-batch log-posterior $\widehat{\mathcal{LM}}_b$
1	−205.7	−14.8	−220.5
2	−205.5	−14.3	−219.8
3	−205.8	−15.5	−221.3
4	−206.0	−14.6	−220.6
5	−205.5	−14.4	−219.9
6	−206.5	−15.3	−221.8
7	−206.2	−14.5	−220.7
8	−205.3	−14.4	−219.7
9	−205.8	−15.0	−220.8
10	−205.4	−15.0	−220.4
11	−207.0	−14.2	−221.2
12	−206.0	−15.2	−221.2
13	−205.4	−13.8	−219.2
14	−205.7	−14.3	−220.0
15	−205.9	−14.5	−220.4

have an approximate t distribution with $(B - 1)$ df. Using this approximation, a 95% confidence interval for the compound Laplace–Metropolis estimate is $(-222.0, -219.0)$. The actual log-integrated likelihood was -221.8. So the compound Laplace–Metropolis estimator worked well in this example.

8. ONE OBSERVATION PER RANDOM EFFECT

The compound Laplace–Metropolis estimator is justified by an asymptotic argument as the number of observations for each random effect becomes large. In practice, the number of observations per random effect may often be small, so it is important to check that the approximation is accurate in this case. In hierarchical models whose component distributions are smooth, unimodal, and not too heavy-tailed, it seems reasonable to expect this to be so, because the integrands will then usually be smooth, bounded unimodal, and fairly light-tailed, even if they are not highly concentrated. In one non-Gaussian example, Grunwald, Raftery, and Guttorp (1993) found the Laplace method to give results accurate to three significant figures, even when the likelihood involved in the integral was based on only one observation.

To see whether the Laplace approximation was adequate for the WFS example of Section 6, we checked the Laplace approximation for the extreme case where there was only one observation for each random effect, over the full range of fitted values encountered. To do this, we calculated conditional likelihoods as in Equation (10) using the Laplace approximation and then using the adaptive 15-point Gauss–Kronrod quadrature method implemented in the S-PLUS integrate function; the latter gives a result that is essentially exact. We performed these calculations for several values of $\mathbf{X}_{it}\tilde{\eta}$, covering the entire range of such values found in the WFS example. We looked at how good the approximation was when there was only a single observation per random effect (i.e., when the ith woman's response, $\mathbf{Y}_i$, was just a scalar 0 or 1) and when there were only two observations per random effect, when $\mathbf{Y}_i \in \{\binom{0}{0}, \binom{0}{1}, \binom{1}{0}, \binom{1}{1}\}$. Table 3 gives the results of this investigation.

As can be seen in Table 3, the Laplace approximation was accurate to at least three digits after the decimal point in all cases. In other words, even when there are only one or two observations per random effect, the Laplace approximation still works very well. As a further check, we calculated the log-conditional likelihood for each woman in the sample used in the example in Section 6 by Gaussian quadrature. We compared these actual log-conditional likelihoods to the log-conditional likelihoods that we previously estimated using the Laplace approximation. We found that the estimated log-conditional likelihoods equaled the results using quadrature to at least three digits after the decimal point. So, at least for the logistic hierarchical model applied in this article, there is no reason to be concerned by using the Laplace approximation in the compound Laplace–Metropolis estimator.

9. DISCUSSION

Raftery (1996a) originally proposed the Laplace–Metropolis estimator to get around limitations encountered when trying to use the Laplace method. He also proposed using the L_1 center as an approximation for the posterior mode in the situation where it is impractical to calculate the likelihood or the log-likelihood for each simulated parameter vector. A reader might wonder why we needed to use it here, because, as we noted in Section 6, we were able to find the "actual" log-integrated likelihood by taking a sample of 50 million draws from the prior distribution. The program to do this took about 5 days on our SPARCstation II. The compound Laplace–Metropolis approximation was found in about 2 hours—a major reduction. Using the compound Laplace–Metropolis estimator also requires noticeably less computer time than the adaptive 15-point Gauss–Kronrod quadrature. Moreover, along with these substantial savings in computer time, the approximations obtained using the compound Laplace–Metropolis estimator are remarkably accurate.

Table 3. Checking the Laplace Approximation With Only One or Two Observations Per Random Effect

$\mathbf{X}_{it}\tilde{\eta}$	$\mathbf{Y}_i = (0)$		$\mathbf{Y}_i = (1)$	
	Laplace	Exact	Laplace	Exact
−4	−.0189	−.0190	−3.9750	−3.9750
−2	−.1310	−.1311	−2.0969	−2.0969
0	−.6933	−.6931	−.6933	−.6932
$\mathbf{X}_{it}\tilde{\eta}$	$\mathbf{Y}_i = \binom{0}{0}$		$\mathbf{Y}_i = \binom{1}{1}$	
	Laplace	Exact	Laplace	Exact
−4	−.0378	−.0379	−7.8625	−7.8625
−2	−.2607	−.2607	−4.1255	−4.1255
0	−1.3649	−1.3647	−1.3649	−1.3647
$\mathbf{X}_{it}\tilde{\eta}$	$\mathbf{Y}_i = \binom{0}{1}$ or $\binom{1}{0}$			
	Laplace	Exact		
−4	−3.9957	−3.9957		
−2	−2.2379	−2.2379		
0	−1.4086	−1.4084		

NOTE: The exact value was calculated using adaptive 15-point Gauss–Kronrod quadrature.

The methods described here should not be expected to work well if the analytic Laplace approximation (4) itself (without posterior simulation) to the overall integrated likelihood is poor, or if the Laplace approximation (6) to the integral over random effects is poor. General conditions for the Laplace approximation to work well have been studied by Kass, Tierney, and Kadane (1988, 1990, 1991), Tierney and Kadane (1986), and Tierney, Kass, and Kadane (1989a,b). One might expect the Laplace approximation in its analytic form—and hence also the simulation-based version used here—to fail if the posterior distribution is highly nonnormal, particularly if it is long-tailed. However, in several empirical studies, including the present one, the Laplace method has been found to be accurate even when the conditions for it to be so are not clearly met (Achcar and Smith 1990; Grunwald et al. 1993; Leonard and Hsu 1994; Raftery 1996b). One would expect the Laplace approximation (6) to the integral over random effects to work well if *either* the prior distribution of the random effects *or* the likelihood (9) for an individual random effect is bounded and dominated by a single mode, without unduly long tails.

As noted previously, the standard Bayesian solution for comparing two models is to compute the Bayes factor. The Bayes factor is the ratio of two integrated likelihoods. In this article we have demonstrated how we were able to obtain good approximations for integrated likelihoods in hierarchical models. We found a point estimate of -220.5 for the log-integrated likelihood for an example model with four fixed effects. We have also fit a number of other models to the IFS event history dataset. For example, when we add husband's level of completed education to the model, we get an approximate log-integrated likelihood of -223.5. Hence the Bayes factor found for comparing the model without husband's education to the model with husband's education is approximately $e^{3.0} \approx 20$, providing evidence for the smaller model. Models containing other potential covariates may be similarly compared.

Bayes factors can be used not only to compare models containing different fixed-effects-parameters, but also to assess whether the data provide evidence for or against the presence of random effects. In Section 7 we found that a compound Laplace–Metropolis estimate of the log-integrated likelihood for the four–fixed-effects model was -220.5. This model included a random-effects parameter for each woman in the sample. Should we have included the random effects in the model? If we can calculate the log-integrated likelihood of a model without random effects, then we will be able to determine a Bayes factor for comparing the model with random effects to the model without random effects.

The integrated likelihood for the model without random effects can be approximated using the Laplace method (Raftery 1996b). For the four–fixed-effects model, this was -222.15. Hence the Bayes factor for comparing the model with random effects to the model without random effects is $\exp\{-220.5 - (-222.15)\} \approx 1.35$. There is some, but only weak, evidence favoring the model incorporating random effects.

REFERENCES

Achcar, J. A., and Smith, A. F. M. (1990), "Aspects of Reparametrization in Approximate Bayesian Inference," in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A Barnard*, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland, pp. 439–452.

Carlin, B. P., and Chib, S. (1993), "Bayesian Model Choice via Markov Chain Monte Carlo," Research Report 93-006, University of Minnesota, Division of Biostatistics.

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

Cleland, J. G., and Scott, C. (Eds.) (1987), *The World Fertility Survey: An Assessment*, Oxford, U.K.: Oxford University Press.

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

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

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473–511.

Green, P. (1995), "Reversible Jump MCMC Computation and Bayesian Model Determination," technical report, University of Bristol, Dept. of Mathematics.

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993), "Time Series of Continuous Proportions," *Journal of the Royal Statistical Society*, Ser. B, 55, 103–116.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

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

Johnson, N. L., and Kotz, S. (1985), *Encyclopedia of Statistical Sciences* (Vol. 5), New York: Wiley.

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

Kass, R. E., Tierney, L., and Kadane, J. B. (1988), "Asymptotics in Bayesian Computation," 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. 261–278.

——— (1990), "The Validity of Posterior Expansions Based on Laplace's Method," in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, Amsterdam: North-Holland, pp. 473–488.

——— (1991), "Laplace's Method in Bayesian Analysis," in *Statistical Multiple Integration*, eds. N. Flournoy and R. K. Tsutakawa, Providence, RI: American Mathematical Society, pp. 89–99.

Leonard, T., and Hsu, J. (1994), "The Laplacian *T*-Approximation in Bayesian Inference," *Statistica Sinica*, 4, 127–142.

Lewis, S. M. (1993), Comment on "Three Papers on Gibbs Sampling and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 79–81.

——— (1994), "Multilevel Modeling of Discrete Event History Data Using Markov Chain Monte Carlo Methods," unpublished Ph.D. dissertation, University of Washington, Dept. of Statistics.

Madigan, D. M., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

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.

Raftery, A. E. (1996a), "Hypothesis Testing and Model Selection," in

Markov Chain Monte Carlo in Practice, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 165–187.

—— (1996b), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, 83, 251–266.

Raftery, A. E., Lewis, S. M., and Aghajanian, A. (1995), "Demand or Ideation? Evidence From the Iranian Marital Fertility Decline," *Demography*, 32, 159–182.

Raftery, A. E., Lewis, S. M., Aghajanian, A., and Kahn, M. J. (1996), "Event History Modeling of World Fertility Survey Data," *Mathematical Population Studies*, 6, 129–153.

Raudenbush, S. W. (1988), "Educational Applications of Hierarchical Linear Models: A Review," *Journal of Educational Statistics*, 13, 85–116.

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), Discussion of "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, 395–402.

Schmeiser, B. (1982), "Batch Size Effects in the Analysis of Simulation Output," *Operations Research*, 30, 556–568.

Terrell, G. R. (1990), "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, 85, 470–477.

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

Tierney, L., Kass, R. E., and Kadane, J. B. (1989a), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716.

—— (1989b), "Approximate Marginal Densities of Nonlinear Functions," *Biometrika*, 76, 425–433.