Marginal Likelihood From the Metropolis–Hastings Output

Siddhartha CHIB and Ivan JELIAZKOV

This article provides a framework for estimating the marginal likelihood for the purpose of Bayesian model comparisons. The approach extends and completes the method presented in Chib (1995) by overcoming the problems associated with the presence of intractable full conditional densities. The proposed method is developed in the context of MCMC chains produced by the Metropolis–Hastings algorithm, whose building blocks are used both for sampling and marginal likelihood estimation, thus economizing on prerun tuning effort and programming. Experiments involving the logit model for binary data, hierarchical random effects model for clustered Gaussian data, Poisson regression model for clustered count data, and the multivariate probit model for correlated binary data, are used to illustrate the performance and implementation of the method. These examples demonstrate that the method is practical and widely applicable.

KEY WORDS: Bayes factor; Bayesian model comparison; Clustered count data; Correlated binary data; Local invariance; Local reversibility; Metropolis–Hastings algorithm; Multivariate density estimation; Reduced conditional density.

1. INTRODUCTION

Consider the problem of comparing a collection of models $\{\mathcal{M}_1, \ldots, \mathcal{M}_L\}$ that reflect competing hypotheses about the data $\mathbf{y} = (y_1, \ldots, y_n)$. Suppose that each model $\mathcal{M}_l$ is characterized by a model-specific parameter vector $\boldsymbol{\theta}_l \in S_l \subseteq \Re^{k_l}$ of dimension k_l and sampling density $f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)$. In this context, Bayesian model selection proceeds by pairwise comparison of the models in $\{\mathcal{M}_l\}$ through their posterior odds ratio, which for any two models $\mathcal{M}_i$ and $\mathcal{M}_j$ is written as

$$\frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \times \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)} \tag{1}$$

where

$$m(\mathbf{y}|\mathcal{M}_l) = \int f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)\, \pi_l(\boldsymbol{\theta}_l|\mathcal{M}_l)\, d\boldsymbol{\theta}_l \tag{2}$$

is the marginal likelihood of $\mathcal{M}_l$. The first fraction on the right-hand side of (1) is known as the *prior odds* and the second as the *Bayes factor*.

The calculation of the marginal likelihood, which is of some importance in Bayesian statistics, has attracted considerable interest in the recent Markov chain Monte Carlo (MCMC) literature. One generic problem is that because the marginal likelihood is obtained by integrating the sampling density $f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)$ with respect to the prior distribution of the parameters, and not the posterior distribution, the posterior MCMC output from the simulation cannot be used directly to estimate the marginal likelihood. Of course, analytic evaluation of the integral is almost never possible. Because of these difficulties, attempts have been made (for example, Carlin and Chib 1995; Green 1995) to estimate the posterior odds by Markov chain Monte Carlo methods that sample both model space and parameter space. These approaches deliver the posterior probabilities of each model, and hence the posterior odds for any two models, according to the frequency of visits to each model. Although such methods are important, they suffer from certain drawbacks. One is that the tuning of the MCMC samplers to promote mixing on a high-dimensional space can be difficult (see Han and Carlin 2000) especially in models with latent variables as in the examples that follow. Even in the best of circumstances, these methods require prerun tuning to get suitable mixing on model space. Another problem is that some subset of the existing models must be included in the simulation if a new model is to be compared to the existing ones, increasing the dimensionality of the parameter space and introducing tuning concerns beyond those required for sampling from the new model.

Work has also been done on the direct estimation of the marginal likelihood in general non-nested model settings (Chib 1995; Gelfand and Dey 1994) and the estimation of ratios of marginal likelihoods especially in the setting of nested models (Chen and Shao 1997; DiCiccio, Kass, Raftery, and Wasserman 1997; Meng and Wong 1996; Verdinelli and Wasserman 1995). In this article, we focus on the approach of Chib (1995), which is based on a representation of the marginal likelihood that is amenable to calculation by MCMC methods. Because the marginal likelihood is the normalizing constant of the posterior density, one can write

$$m(\mathbf{y}|\mathcal{M}_l) = \frac{f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)\pi(\boldsymbol{\theta}_l|\mathcal{M}_l)}{\pi(\boldsymbol{\theta}_l|\mathbf{y}, \mathcal{M}_l)}, \tag{3}$$

which is referred to as the *basic marginal likelihood identity*. Evaluating the right-hand side of this identity at some appropriate point $\boldsymbol{\theta}_l^*$ and taking logarithms one obtains the expression

$$\log m(\mathbf{y}|\mathcal{M}_l) = \log f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l^*) + \log \pi(\boldsymbol{\theta}_l^*|\mathcal{M}_l) - \log \pi(\boldsymbol{\theta}_l^*|\mathbf{y}, \mathcal{M}_l) \tag{4}$$

from which the marginal likelihood can be estimated by finding an estimate of the posterior ordinate $\pi(\boldsymbol{\theta}_l^*|\mathbf{y}, \mathcal{M}_l)$. Thus the calculation of the marginal likelihood is reduced to finding an estimate of the posterior density at a single point $\boldsymbol{\theta}_l^*$. For estimation efficiency, the latter point is generally taken to be a high-density point in the support of the posterior.

Siddhartha Chib is Professor of Econometrics and Statistics, John M. Olin School of Business, CB1133, Washington University, St. Louis, MO 63130 (Email: *chib@olin.wustl.edu*). Ivan Jeliazkov is a doctoral student in Economics at Washington University. The authors are grateful to the editor, associate editor, and referees for helpful comments.

Chib (1995) provides a method to estimate the posterior ordinate in the context of Gibbs MCMC sampling. Suppose that the parameter space is split into B conveniently specified blocks, so that $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_B^*)$ where we suppress the model index for notational convenience. Then, by the law of total probability we have

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \pi(\boldsymbol{\theta}_1^*|\mathbf{y})\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)\cdots\pi(\boldsymbol{\theta}_B^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{B-1}^*) \quad (5)$$

where $\pi(\boldsymbol{\theta}_1^*|\mathbf{y})$ is the marginal density ordinate of $\boldsymbol{\theta}_1$ and $\pi(\boldsymbol{\theta}_B^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{B-1}^*)$ is the full conditional density ordinate and the remaining ordinates are reduced conditional ordinates. Now assume that each full conditional density is fully known. Then, the marginal density ordinate is estimated by the Rao–Blackwell device (Gelfand and Smith 1990; Tanner and Wong 1987). Next, the first reduced conditional ordinate is found by averaging the full conditional density of $\boldsymbol{\theta}_2$, $\hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) = M^{-1}\sum_{j=1}^{M}\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_B^{(j)})$ where $\{\boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_B^{(j)}\} \sim \pi(\boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_B|\mathbf{y}, \boldsymbol{\theta}_1^*)$ are M draws that are obtained from a *reduced Gibbs* MCMC *run* in which $\boldsymbol{\theta}_1$ is fixed at $\boldsymbol{\theta}_1^*$ and sampling is over $\{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B\}$, a procedure that requires no new programming. Subsequent reduced ordinates are estimated in the same way by fixing additional blocks. The time cost of this procedure is generally small when blocking is done effectively and a few reduced runs are required, as is possible in many practical problems, and the higher-dimensional blocks are placed first in the equation (5) decomposition.

It is worth noting that an alternative approach to estimating the posterior ordinate is developed by Ritter and Tanner (1992), also in the context of Gibbs MCMC chains with fully known full conditional distributions. If one lets

$$K_{\mathrm{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y}) = \prod_{k=1}^{B}\pi(\boldsymbol{\theta}_k^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k-1}^*, \boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_B) \quad (6)$$

denote the Gibbs transition kernel, then by virtue of the fact that the Gibbs chain satisfies the invariance condition $\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \int K_{\mathrm{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}$, one can obtain the posterior ordinate by averaging the transition kernel over draws from the posterior distribution $\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) = M^{-1}\sum_{g=1}^{M}K_{\mathrm{G}}(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^*|\mathbf{y})$. This estimate only requires draws from the full Gibbs run, but when $\boldsymbol{\theta}$ is high-dimensional and the model contains latent variables, this estimate tends to be less accurate than Chib's estimate, which achieves accuracy with the help of additional simulations by decomposing the posterior ordinate into smaller pieces and estimating each reduced ordinate from its full conditional distribution. Nonetheless, the idea embodied in Ritter and Tanner's method is valuable, and it is a useful method when $\boldsymbol{\theta}$ is low-dimensional.

Neither the Chib nor Ritter and Tanner methods for estimating the posterior ordinate offer a solution to problems in which sampling is via a single-block Metropolis sampler or for problems in which one or more of the normalizing constants of the full conditional densities are not known. For the latter situation, Chib and Greenberg (1998) have applied a nonparametric density estimation method to calculate the recalcitrant reduced ordinate, but in a high-dimensional case the accuracy of the resulting estimate is questionable, although they suggest a modified method that Chib, Nardari, and Shephard (1999) have applied to a 120-dimensional posterior distribution. In this article, we propose a different, less demanding, and more general solution that solves both problems, thus extending and completing the Chib method for estimating the marginal likelihood.

The rest of the article is organized as follows. In Section 2, we discuss the MCMC sampling framework and give the main results that form the basis for the proposed method. Section 3 contains results from extensive experiments that document the performance of the method in diverse model situations. Concluding remarks are given in Section 4.

2. PROPOSED APPROACH

2.1 One Block Sampling

To motivate the general approach, we begin by considering a simple case of some importance. Suppose that the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})$, defined over S, a subset of $\Re^d$, is sampled in one block by the Metropolis–Hastings algorithm (Chib and Greenberg 1995; Hastings 1970; Tierney 1994) and the goal is to estimate the posterior ordinate $\pi(\boldsymbol{\theta}^*|\mathbf{y})$ given the posterior sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\}$. Let $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ denote the proposal density (candidate generating density) for the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, where the proposal density is allowed to depend on the data $\mathbf{y}$, and let

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})}\right\}$$

denote the probability of move (probability of accepting the proposed value).

If we let $p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ denote the subkernel of the M–H algorithm, then from the reversibility of the subkernel we can write that for any point $\boldsymbol{\theta}^*$

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}^*|\mathbf{y})p(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y}). \quad (7)$$

Upon integrating both sides of this expression with respect to $\boldsymbol{\theta}$ over $\Re^d$, we immediately obtain the result that the posterior ordinate is given by

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{\int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}}. \quad (8)$$

To highlight the estimation strategy, write the latter in more succinct form as

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{E_1\{\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{y})\}}{E_2\{\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})\}}$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ and the denominator expectation E_2 is with respect to $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})$. This implies that a simulation-consistent estimate of the posterior ordinate is available as

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^*|\mathbf{y})}{J^{-1}\sum_{j=1}^{J}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}|\mathbf{y})}. \quad (9)$$

where $\{\boldsymbol{\theta}^{(g)}\}$ are the sampled draws from the posterior distribution and $\{\boldsymbol{\theta}^{(j)}\}$ are draws from $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})$, given the fixed

value $\boldsymbol{\theta}^*$. On substituting the latter estimate in the log of the basic marginal likelihood identity, we get

$$\log \hat{m}(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) \qquad (10)$$

This is a rather simple expression that can be applied to many models including the class of univariate generalized linear models, parametric survival models, and Gaussian linear mixed models for clustered data. Note that the choice of point $\boldsymbol{\theta}^*$ is arbitrary, but for estimation efficiency it is customary to choose a point that has high density under the posterior. Also note that we let the simulation-sample sizes in the numerator and the denominator be different, although in practice we set them to be equal. It should be appreciated that the sampling of $\boldsymbol{\theta}^{(j)}$ from $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})$ normally consumes a small amount of time in relation to the time required for the full MCMC run. This means that the marginal likelihood of the model is available almost as soon as the full MCMC run is finished. Finally, if S is a proper subset of $\Re^d$, then values $\boldsymbol{\theta}^{(j)}$ from $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{y})$ that do not lie in S are included in the average in the denominator with the value $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}|\mathbf{y}) = 0$.

2.2 Two Parameter Blocks and Multiple Latent Variable Blocks

To enhance understanding, assume that the normalizing constant of only the first full conditional density is not known and that this density is sampled by the M–H algorithm, as in the multivariate probit model example presented in section 3.5. Let $q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$ denote the proposal density for the transition from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_1'$ where we have for generality allowed the proposal density to depend on the data and the two remaining blocks. Also let

$$\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}) = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}_1', \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2)}{f(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \times \frac{q(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})}{q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1'|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})}\right\}$$

denote the probability of move. An application of the basic marginal likelihood identity yields

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)}{\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*|\mathbf{y})}$$

and the goal is to estimate $\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*|\mathbf{y})$. We assume that the likelihood ordinate is available readily either by direct calculation (as in mixed effects models for clustered data where $\mathbf{z}$ denotes the cluster-specific random effects) or by a Monte Carlo integration method. Note that although it is also true that $m(\mathbf{y}) = f(\mathbf{y}, \mathbf{z}^*|\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)/\pi(\mathbf{z}^*, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*|\mathbf{y})$, the latter identity is not that useful because it requires the computation of the ordinate $\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{z}^*|\mathbf{y})$ whose dimension can easily run into the hundreds, if not thousands.

Following Chib (1995), we decompose the posterior ordinate as

$$\pi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*|\mathbf{y}) = \pi(\boldsymbol{\theta}_1^*|\mathbf{y})\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*),$$

where $\pi(\boldsymbol{\theta}_1^*|\mathbf{y})$ cannot be estimated by the Rao–Blackwell method because, by assumption, the normalizing constant of $\pi(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$ is not known. Let

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}) = \alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$$

denote the subkernel of the M–H chain for $\boldsymbol{\theta}_1$ conditioned on $(\boldsymbol{\theta}_2, \mathbf{z})$. Then, by direct calculation we see that this kernel satisfies the condition

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}) = \pi(\boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z}), \qquad (11)$$

which may be referred to as a *local reversibility condition*. Now if we multiply both sides of (11) by $\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})$ and integrate over $\boldsymbol{\psi} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z})$, we get

$$\int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})\, d\boldsymbol{\psi} = \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})\, d\boldsymbol{\psi}$$

or

$$\int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})\, d\boldsymbol{\psi} = \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1^*|\mathbf{y})\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)\, d\boldsymbol{\psi}$$

and so,

$$\int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})d\boldsymbol{\psi} = \pi(\boldsymbol{\theta}_1^*|\mathbf{y})\int p(\theta_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)d\boldsymbol{\psi}.$$

From this last equality, and the definitions of $p(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$ and $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$, it follows that

$$\pi(\boldsymbol{\theta}_1^*|\mathbf{y}) = \frac{E_1\{\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\}}{E_2\{\alpha(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})\}}, \qquad (12)$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y})$, whereas the denominator expectation E_2 is with respect to the distribution $\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*) \times q(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$.

Each of the integrals in (12) can be estimated by the Monte Carlo method. To estimate the numerator, we take the draws $\{\boldsymbol{\theta}_1^{(g)}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)}\}_{g=1}^M$ from the full run and average the quantity $\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$. For the denominator, because the expectation is conditioned on $\boldsymbol{\theta}_1^*$, we continue the MCMC simulation for an additional J iterations with the two full conditional densities

$$\pi(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z}); \qquad \pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2) \qquad (13)$$

At each iteration of this reduced run, given the values $(\boldsymbol{\theta}_2^{(j)}, \mathbf{z}^{(j)})$, we generate a variate

$$\boldsymbol{\theta}_1^{(j)} \sim q(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(j)}, \mathbf{z}^{(j)})$$

leading to the triple $(\boldsymbol{\theta}_2^{(j)}, \mathbf{z}^{(j)}, \boldsymbol{\theta}_1^{(j)})$ that is a draw from the distribution $\pi(\boldsymbol{\theta}_2, \mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)q(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$. The marginal ordinate can now be estimated as

$$\hat{\pi}(\boldsymbol{\theta}_1^*|\mathbf{y}) = \frac{M^{-1}\sum_{g=1}^M \alpha(\boldsymbol{\theta}_1^{(g)}, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)})q(\boldsymbol{\theta}_1^{(g)}, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)})}{J^{-1}\sum_{j=1}^J \alpha(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1^{(j)}|\mathbf{y}, \boldsymbol{\theta}_2^{(j)}, \mathbf{z}^{(j)})}. \qquad (14)$$

Next, the values $\mathbf{z}^{(j)}$ from the preceeding reduced run, which are marginally from $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)$, are used to form the average $\hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) = J^{-1}\sum_{j=1}^{J} \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z}^{(j)})$, which is a simulation-consistent estimate of $\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)$. Thus at the conclusion of the reduced run, both ordinates are available and the marginal likelihood estimate is given by

$$\log \hat{m}(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \{\log \hat{\pi}(\boldsymbol{\theta}_1^*|\mathbf{y}) + \log \hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)\}. \quad (15)$$

Three remarks are in order. First, a single reduced run, augmented with a step to sample $\boldsymbol{\theta}_1$ from the proposal density, delivers the variates that are used in the calculation of each of the two ordinates, $\pi(\boldsymbol{\theta}_1^*|\mathbf{y})$ and $\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)$. Second, if one places the recalcitrant ordinate first in the decomposition of the posterior ordinate, then the reduced run does not involve any M–H steps. Third, as mentioned previously and shown later in Section 3.4, this same approach can be applied when the full conditional density of $\mathbf{z}$ is sampled by a sequence of M–H steps.

2.3 Multiple Parameter Blocks

One of the features of the approach of Chib (1995), which has been inherited by the current method, is the flexibility with which it accommodates both low- and high-dimensional problems. For example, in most low-dimensional MCMC problems, grouping all parameters into one block is a sensible strategy. In higher-dimensional problems, it may be the case that convenience, computational necessity, or simulation design may require sampling of the parameters in several smaller, more manageable blocks. Our approach readily generalizes to this situation.

To describe the setting, suppose that the MCMC sampling is conducted without any latent data, because latent data, sampled in one or more blocks, can be handled as in the previous section, and suppose that the parameters are grouped into B blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B)$, with $\boldsymbol{\theta}_k \in S_k \subseteq \Re^{d_k}$. Write the posterior ordinate at $\boldsymbol{\theta}^*$ as $\pi(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_B^*|\mathbf{y}) = \prod_{i=1}^{B} \pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*)$ and consider the estimation of the reduced ordinate $\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*)$.

Now suppose that the full conditional density $\pi(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$, $i = 1, \ldots, B$, is sampled by the M–H algorithm with proposal density $q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i'|\mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1})$ and probability of move

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i'|, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1}) = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}_i', \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1})\pi(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{-i})}{f(\mathbf{y}|\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1})\pi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i})} \times \frac{q(\boldsymbol{\theta}_i', \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1})}{q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i'|\mathbf{y}, \boldsymbol{\psi}_{i-1}, \boldsymbol{\psi}^{i+1})}\right\},$$

where we have written $\boldsymbol{\psi}_{i-1} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1})$ to denote the blocks up to i and $\boldsymbol{\psi}^{i+1} = (\boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_B)$ to denote the blocks beyond i.

Again by exploiting the local reversibility of the M–H step for $\boldsymbol{\theta}_i$ and completely analogous arguments to the ones presented in the last subsection, we obtain the result that

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*) = \frac{E_1\{\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}}{E_2\{\alpha(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}} \quad (16)$$

where E_1 is the expectation with respect to conditional posterior $\pi(\boldsymbol{\theta}_i, \boldsymbol{\psi}^{i+1}|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*)$ and E_2 that with respect to the conditional product measure $\pi(\boldsymbol{\psi}^{i+1}|\mathbf{y}, \boldsymbol{\psi}_i^*)q(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})$. These two integrals can be estimated as before from the output of the reduced MCMC runs, as follows.

Step 1. Set $\boldsymbol{\psi}_{i-1} = \boldsymbol{\psi}_{i-1}^*$ and sample the reduced set of full conditional distributions $\pi(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_{-k})$, $k = i, \ldots, B$. Let the generated draws be $\{\boldsymbol{\theta}_i^{(g)}, \ldots, \boldsymbol{\theta}_B^{(g)}\}$, $g = 1, \ldots, M$.

Step 2. Include $\boldsymbol{\theta}_i^*$ in the conditioning set, let $\boldsymbol{\psi}_i^* = (\boldsymbol{\psi}_{i-1}^*, \boldsymbol{\theta}_i^*)$ and remove the $\boldsymbol{\theta}_i$ full conditional distribution from the collection in Step 1. Then sample the remaining distributions $\pi(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_{-k})$, $k = i+1, \ldots, B$, to produce $\{\boldsymbol{\theta}_{i+1}^{(j)}, \ldots, \boldsymbol{\theta}_B^{(j)}\}$. At each step of the sampling also draw $\boldsymbol{\theta}_i^{(j)}$ from $q(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1,(j)})$.

Step 3. Estimate the reduced ordinate in (16) by the ratio of Monte Carlo averages

$$\hat{\pi}(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*) = \frac{M^{-1}\sum_{g=1}^{M} \alpha(\boldsymbol{\theta}_i^{(g)}, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1,(g)})q(\boldsymbol{\theta}_i^{(g)}, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1,(g)})}{J^{-1}\sum_{j=1}^{J} \alpha(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i^{(j)}|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1,(j)})} \quad (17)$$

where the average in the denominator may include zeros if there are $\boldsymbol{\theta}_i^{(j)}$ values that lie outside the support of the posterior S.

Step 4. Estimate the marginal likelihood on the log scale as

$$\log \hat{m}(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \sum_{i=1}^{B} \log \hat{\pi}(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*) \quad (18)$$

It is important to keep in mind that the reduced runs are obtained by fixing an appropriate set of parameters and continuing the MCMC simulation with a smaller set of distributions. Therefore, these runs require little to no coding beyond what is done initially for the sampling of the posterior distribution. As long as the full MCMC sampling scheme has been properly designed (with blocking and proposal densities that avoid reducibility problems), each of the reduced runs is well defined. Finally, observe that the variates $\boldsymbol{\psi}^{(i+1)}$ in Step 2 are automatically produced in the next reduced run, where the ordinate $\pi(\boldsymbol{\theta}_{i+1}^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_i^*)$ is estimated.

2.4 Numerical Standard Error of the Marginal Likelihood Estimate

In this section, we discuss briefly how the numerical standard error of the marginal likelihood estimate can be derived. The numerical standard error gives the variation that can be expected in the marginal likelihood estimate if the simulation were to be repeated. For specificity, we consider the case in Section 2.2 under the assumption that $M = J$. Following Chib (1995) we define the vector process

$$\mathbf{h}^{(g)} = \begin{pmatrix} h_1^{(g)} \\ h_2^{(g)} \\ h_3^{(g)} \end{pmatrix} \equiv \begin{pmatrix} \alpha(\boldsymbol{\theta}_1^{(g)}, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)}) q_1(\boldsymbol{\theta}_1^{(g)}, \boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)}) \\ \alpha(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_1^{(g)}|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)}) \\ \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z}^{(g)}) \end{pmatrix},$$

where the draws in the first component of h are from the full MCMC run, while those in the second and third components are from the reduced run, although this is not emphasized in the notation. If we let $\hat{\mathbf{h}} = M^{-1} \sum_{g=1}^{M} \mathbf{h}^{(g)}$, then by equations (14) and (15), an estimate of the log-marginal likelihood as a function of the elements of $\hat{\mathbf{h}}$ is given by

$$\log \hat{m}(\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \{\log \hat{h}_1 - \log \hat{h}_2 + \log \hat{h}_3\}.$$

To estimate the variance of this quantity, suppose that the values of $f(\mathbf{y}|\boldsymbol{\theta}^*)$ and $\pi(\boldsymbol{\theta}^*)$ are available directly. Then, the variance of the log-marginal likelihood estimate becomes $\operatorname{var}(\log \hat{m}(\mathbf{y})) = \operatorname{var}(\log \hat{h}_1 - \log \hat{h}_2 + \log \hat{h}_3)$, which can be found by the Delta method once we obtain an estimate of the variance of $\hat{\mathbf{h}}$. As $\{\mathbf{h}^{(g)}\}$ inherits the ergodicity of the Markov chain, it follows by the ergodic theorem in Tierney (1994) that $\hat{\mathbf{h}}$ will converge to its mean almost surely as $M \to \infty$. Under regularity conditions, an estimate of the sample variance of $\hat{\mathbf{h}}$ is given by the expression (Newey and West 1987)

$$\operatorname{var}(\hat{\mathbf{h}}) = M^{-1} \left[\boldsymbol{\Omega}_0 + \sum_{s=1}^{m} \left(1 - \frac{s}{m+1}\right) (\boldsymbol{\Omega}_s + \boldsymbol{\Omega}_s') \right],$$

where

$$\boldsymbol{\Omega}_s = M^{-1} \sum_{g=s+1}^{M} (\mathbf{h}^{(g)} - \hat{\mathbf{h}})(\mathbf{h}^{(g-s)} - \hat{\mathbf{h}})', \quad s = 0, 1, \ldots, m,$$

and m is a constant that represents the lag at which the autocorrelation function of $\mathbf{h}^{(g)}$ tapers off. In the examples that follow, we have set the value of m, somewhat cautiously, to equal 40. Given this covariance matrix, it follows from the Delta method that $\operatorname{var}(\log \hat{m}(\mathbf{y})) = \mathbf{a}' \operatorname{var}(\hat{\mathbf{h}}) \mathbf{a}$, where $\mathbf{a} = (\hat{h}_1^{-1}, -\hat{h}_2^{-1}, \hat{h}_3^{-1})'$. The numerical standard error is the square root of this quantity.

We note that if one also needs to estimate the ordinate of the likelihood function, that of the prior, or both, as in the examples presented in Sections 3.3 and 3.4, then the variance of the latter estimates must be incorporated by a separate calculation. In addition, extending the calculation of the numerical standard error to cases involving more than two blocks is straightforward. One proceeds by including (with appropriate arguments and conditioning sets) more elements like h_1 and h_2 for each block that is simulated by Metropolis–Hastings, and elements similar to h_3 for blocks that are sampled directly.

In closing, we mention that in the subsequent examples we have performed a frequency analysis to verify the accuracy of the numerical standard error estimates obtained by the approach described in this section. The posterior simulations are repeated 50 times for each combination of M and J. The standard deviations of the marginal likelihood estimates obtained from these replications are found to closely mirror those from the above approach, thus providing a useful validation of this method.

3. EXAMPLES

We now discuss the performance of the marginal likelihood estimation method by varying different aspects of the MCMC design in several important models. We report evidence on different facets of the design including the choice of the proposal density, the type of blocking and size of blocks, and the sample sizes M and J used to estimate the posterior ordinate. To measure the efficiency of the MCMC parameter sampling scheme, we use the measures $[1 + 2\sum_{k=1}^{\infty} \rho_k(l)]$, where $\rho_k(l)$ is the sample autocorrelation at lag l for the kth parameter in the sampling with the summation truncated according to (say) the Parzen window. The latter quantity is called the *inefficiency factor* or the *autocorrelation time* and may be interpreted as the ratio of the numerical variance of the posterior mean from the MCMC chain to the variance of the posterior mean from hypothetical independent draws. Our conclusion is that the scheme that is efficient for sampling the posterior distribution, as measured by the inefficiency factors, is also efficient for estimating the marginal likelihood. Thus algorithms that have higher inefficiency factors tend to produce marginal likelihood estimates with larger numerical standard errors although they have minor to no significant effect on the point estimate of the marginal likelihood. Conversely, algorithms that are designed to be efficient for the simulation of the parameters are also efficient for the estimation of the marginal likelihood.

3.1 Binary Data Logit Model

To begin, we consider the marginal likelihood computation in the binary data logit model with emphasis on the effect of the proposal density on the marginal likelihood estimate. For this model it is possible to sample the posterior distribution in one block, which allows us to isolate the role of the proposal density without having to consider the influence of the blocking design.

We compare two canonical proposal densities in this setting. The first is a tailored proposal density, and the second is a random walk proposal density. The *tailored proposal density* is defined by letting $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}'|\mathbf{y}) = f_T(\boldsymbol{\theta}'|\mathbf{m}, \mathbf{V}, \nu)$, where $\mathbf{m}$ is the mode of the log target density and $\mathbf{V}$ is the inverse of the negative Hessian of the log-target evaluated at $\mathbf{m}$,

and $f_T(\cdot|\cdot)$ denotes a multivariate-t density with mean $\mathbf{m}$, variance $\nu\mathbf{V}/(\nu-2)$, and ν degrees of freedom. The *random walk proposal density* is defined as $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = f_T(\boldsymbol{\theta}'|\boldsymbol{\theta}, \tau\mathbf{V}, \nu)$ where we use the tailored matrix $\mathbf{V}$, along with a tuning parameter τ, as the dispersion matrix of the proposal.

The model we fit utilizes an $n = 200$ observation data set from Mroz (1987) that deals with the factors that influence participation of the female spouse in the labor market. Seven covariates, non-wife income, number of years of education, years of experience, experience squared, age, number of children less than 6 years old in the household, and number of children greater than 6 years old in the household are used to explain the binary indicator of labor market participation. Including the constant, the full model contains eight parameters. If we assume that the parameters follow the multivariate normal distribution $\phi_8(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0)$, then the posterior distribution of the parameters is given by

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \phi_8(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0) \prod_{i=1}^{200} p_i^{y_i} (1-p_i)^{(1-y_i)}$$

$$p_i = (1+ e^{-\mathbf{x}_i\boldsymbol{\beta}})^{-1}$$

where $y_i \in \{0, 1\}$ is the binary outcome variable and $\mathbf{x}_i$ is the covariate vector on the ith subject in the sample.

In Figure 1, we see that the MCMC simulation of the posterior distribution with the tailored proposal density produces inefficiency factors that are about five times smaller than those from the random walk proposal density for each value of M that is used in the experiment. Next, we consider the estimates of the marginal likelihoods under the two proposal densities for different values of M and J. These results are shown in Table 1, where we find that the marginal likelihood estimate under the tailored proposal density stabilizes up to the second decimal when the sample sizes are $M = 5{,}000$ and $J = 5{,}000$. Further increases in the sample sizes affect the estimate only in the third decimal place. On the other hand, the estimate from the more volatile random walk chain agrees with the tailored estimate up to the first decimal place only when the sample size is about 20,000. In terms of the variability of the estimates, we see that the numerical standard error of the marginal likelihood estimate, which we report in parentheses below each estimate, is about 10 times smaller from the tailored chain in comparison with the random walk chain.

3.2 Longitudinal Hierarchical Model for Clustered Data

In this example, we illustrate the impact of the MCMC blocking scheme on the marginal likelihood estimate. We use two schemes to sample the posterior distribution of the parameters. In the first scheme, multiple blocks are used to sample the posterior distribution. In the second scheme, the posterior samples are drawn in one block, marginalized over the latent data and a subset of the parameters. Within each scheme we

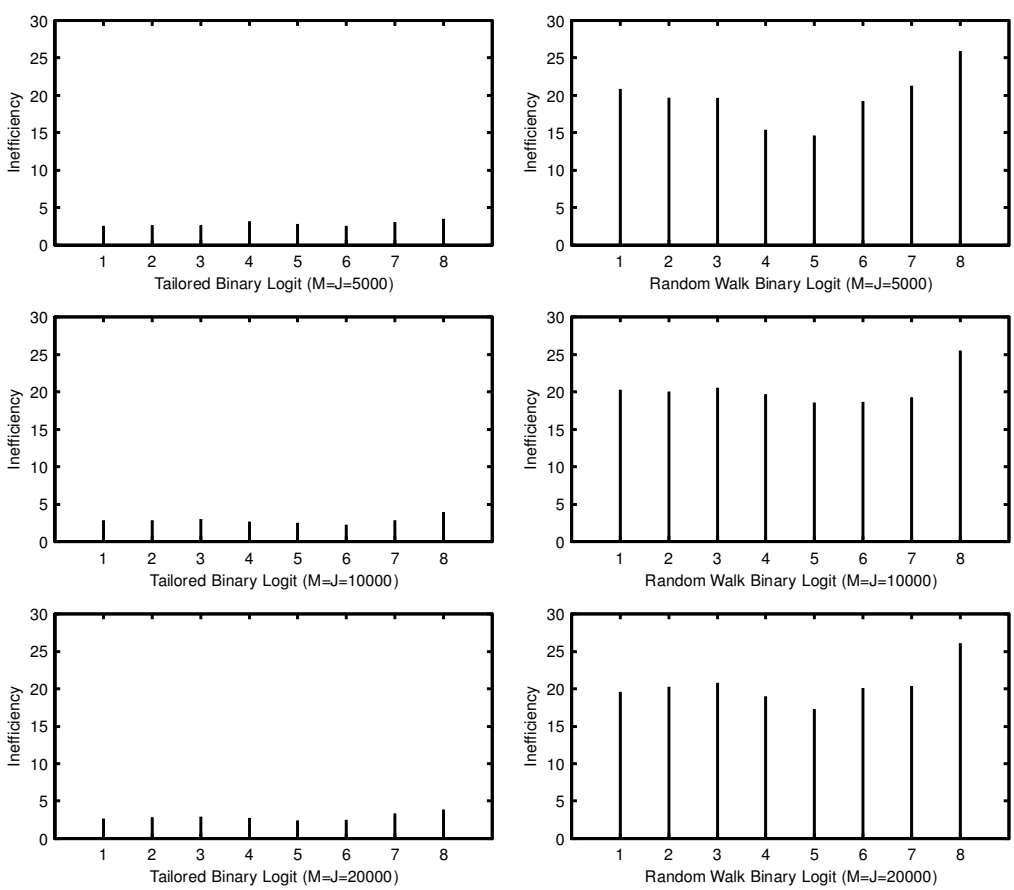

Figure 1. Binary Logit Model. Inefficiency factors under two different proposals and choices of (M,J) for eight covariate parameters.

Table 1. Log-marginal Likelihood Estimates for the Binary Logit Model Using Tailored and Random Walk Proposal Densities

	Type of proposal density	
(M, J)	Tailored	Random walk
(5,000, 5,000)	−144.756 (.015)	−144.835 (.201)
(10,000, 10,000)	−144.757 (.011)	−144.884 (.106)
(20,000, 20,000)	−144.751 (.010)	−144.779 (.103)

are careful to ensure that the sampling is done as efficiently as possible in order to isolate the pure effect of blocking. Otherwise, of course, the deficiencies of the sampling scheme would be confounded with the effect of blocking. Once again, we find that the MCMC parameter simulation scheme that produces the lower inefficiency factors also tends to produce marginal likelihood estimates with lower numerical standard errors.

The effect of blocking is considered in the context of a Gaussian longitudinal data model with random effects. We use the same model and data as Han and Carlin (2000) where several model space methods, including those of Carlin and Chib (1995), and Green (1995), are compared. Han and Carlin report that the model space methods were difficult to implement in this setting.

The data used in our illustration is from a clinical trial on the effectiveness of two antiretroviral drugs (didanosine (ddI) and zalcitabine (ddC)) in 467 persons with advanced HIV infection. The response variable y_{ij} for patient i at time j is the square root of the patient's CD4 count, a seriological measure of immune system health and prognostic factor for AIDS-related illness and mortality. The dataset records patient CD4 counts at study entry and again at 2, 6, 12, and 18 months after entry, for the ddI and ddC groups, respectively. Following the aforementioned work, we fit a mixed-effects model

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\\ \boldsymbol{\varepsilon}_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 I_{n_i})\\ \mathbf{b}_i &\sim \mathcal{N}_2(\mathbf{0}, \mathbf{D})\end{aligned}$$

where the jth row of patient i's design matrix $\mathbf{W}_i$ takes the form $w_{ij} = (1, t_{ij})$, where $t_{ij} \in \{0, 2, 6, 12, 18\}$ and the fixed design matrix $\mathbf{X}_i$ is $\mathbf{X}_i = \left(\mathbf{W}_i | d_i\mathbf{W}_i | a_i\mathbf{W}_i\right)$. The d_i is a binary variable indicating whether patient i received ddI ($d_i = 1$) or ddC ($d_i = 0$), and a_i is a binary variable indicating if the patient was diagnosed as having AIDS at baseline ($a_i = 1$) or not ($a_i = 0$). The prior distribution of $\boldsymbol{\beta} : 6 \times 1$ is assumed to be $\mathcal{N}_6(\boldsymbol{\beta}_0, \mathbf{B}_0)$ with

$$\boldsymbol{\beta}_0 = (10, 0, 0, 0, -3, 0),$$

and

$$\mathbf{B}_0 = \mathrm{Diag}(2^2, 1^2, (.1)^2, 1^2, 1^2, 1^2),$$

that on $\mathbf{D}^{-1}$ is Wishart $\mathcal{W}_2(\rho_0, \mathbf{R}_0/\rho_0)$ with $\rho_0 = 24$ and $\mathbf{R}_0 = \mathrm{diag}(.25, 16)$ and finally that on σ^2 is inverse gamma $\mathcal{IG}(\nu_0/2, \delta_0/2)$ with $\nu_0 = 6$ and $\delta_0 = 400$.

Two blocking schemes are considered for this model both of which rely on the facts, exploited by Chib and Carlin (1999), that

$$\begin{aligned}\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}, \sigma^2 &\sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i)\\ \boldsymbol{\beta}|\mathbf{y}, \mathbf{D}, \sigma^2 &\sim \mathcal{N}_6(\hat{\boldsymbol{\beta}}, \mathbf{B}_n)\end{aligned}$$

where $\boldsymbol{\Omega}_i = (\sigma^2 I_{n_i} + \mathbf{W}_i\mathbf{D}\mathbf{W}_i')$, $\hat{\boldsymbol{\beta}} = \mathbf{B}_n(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{467}\mathbf{X}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{y}_i)$ and $\mathbf{B}_n = (\mathbf{B}_0^{-1} + \sum_{i=1}^{467}\mathbf{X}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i)^{-1}$. Therefore, one MCMC scheme, which we refer to as the multiple-block scheme, is given as follows.

Longitudinal Data Model: Multiple Block MCMC Algorithm

1. Sample $\boldsymbol{\beta} \sim \mathcal{N}_6(\hat{\boldsymbol{\beta}}, \mathbf{B}_n)$ and

$$\mathbf{b}_i \sim \mathcal{N}_2\left\{\mathbf{D}_i(\sigma^{-2}\mathbf{W}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \quad \mathbf{D}_i = (\mathbf{D} + \sigma^{-2}\mathbf{W}_i'\mathbf{W}_i)^{-1}\right\}, \quad i \le 467$$

2. Sample

$$\mathbf{D}^{-1} \sim \mathcal{W}_2\left\{\rho_0 + 467, \left(\mathbf{R}_0^{-1} + \sum_{i=1}^{467}\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\right\}$$

3. Sample

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \sum_{i=1}^{467} \| \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i \|^2}{2}\right).$$

Because each of these distributions is tractable the posterior ordinate can be computed by the direct Chib method using the decomposition

$$\begin{aligned}&\pi(\mathbf{D}^{-1*}, \sigma^{2*}, \boldsymbol{\beta}^*|\mathbf{y})\\ &\quad = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\sigma^{2*}|\mathbf{y}, \mathbf{D}^*)\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^{-1*}, \sigma^{2*}),\end{aligned}$$

where the first term is obtained by averaging the Wishart density over draws on $\{\mathbf{b}_i\}$ from the full run, the second ordinate is estimated by averaging the inverse gamma full density of σ^2 over draws on $(\boldsymbol{\beta}, \{\mathbf{b}_i\})$ from a reduced run conditioned on $\mathbf{D}^*$, and the third ordinate is multivariate normal as given previously and available directly.

In the second scheme, the parameters of the model are sampled in one block by using the fact that the density of the observations marginalized over both $(\boldsymbol{\beta}, \{\mathbf{b}_i\})$ can be expressed as

$$f(\mathbf{y}|\mathbf{D}, \sigma^2) = \frac{\phi_6(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_0, \mathbf{B}_0)\Pi_{i=1}^{467}\phi_{n_i}(\mathbf{y}_i|\mathbf{X}_i\hat{\boldsymbol{\beta}}, \boldsymbol{\Omega}_i)}{\phi_6(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\beta}}, \mathbf{B}_n)},$$

which is an application of the basic marginal likelihood identity, now used to find the marginal density of $\mathbf{y}$ conditioned on $(\mathbf{D}, \sigma^2)$. Therefore, the posterior of $(\mathbf{D}^{-1}, \sigma^2)$ is proportional to this density times the prior densities on these parameters and can be sampled by a one-block tailored M–H method. In particular, let $f_{\mathrm{W}}(\mathbf{D}^{-1}|\rho_0, \mathbf{R}_0)$ and $f_{\mathrm{IG}}(\sigma^2|\nu_0/2, \delta_0/2)$ denote the Wishart and inverse gamma prior densities, respectively,

and let $(\mathbf{m}, \mathbf{V})$ denote the modal value and inverse of the negative Hessian at the mode of the function $\log f(\mathbf{y}|\mathbf{D}^{-1}, \sigma^2) + \log f_{\mathrm{W}}(\mathbf{D}^{-1}|\rho_0, \mathbf{R}_0) + \log f_{\mathrm{IG}}(\sigma^2|\nu_0/2, \delta_0/2)$ and let $q(\mathbf{D}^{-1}, \sigma^2|\mathbf{y}) = f_T(\mathbf{D}^{-1}, \sigma^2|\mathbf{m}, \mathbf{V}, \nu)$ denote the proposal density. Then, in the one block algorithm for the clustered Gaussian model we propose $(\mathbf{D}^{-1'}, \sigma^{2'}) \sim f_T(\mathbf{D}^{-1}, \sigma^2|\mathbf{m}, \mathbf{V}, \nu)$ and move to $(\mathbf{D}^{-1'}, \sigma^{2'})$ with probability

$$\alpha = \min\left\{ \frac{f(\mathbf{y}|\mathbf{D}^{-1'}, \sigma^{2'}) f_{\mathrm{W}}(\mathbf{D}^{-1'}|\rho_0, \mathbf{R}_0) f_{\mathrm{IG}}(\sigma^{2'}|\nu_0/2, \delta_0/2)}{f(\mathbf{y}|\mathbf{D}^{-1}, \sigma^2) f_{\mathrm{W}}(\mathbf{D}^{-1}|\rho_0, \mathbf{R}_0) f_{\mathrm{IG}}(\sigma^2|\nu_0/2, \delta_0/2)} \times \frac{f_T(\mathbf{D}^{-1}, \sigma^2|\mathbf{m}, \mathbf{V}, \nu)}{f_T(\mathbf{D}^{-1'}, \sigma^{2'}|\mathbf{m}, \mathbf{V}, \nu)}, 1 \right\}$$

Given the output from this scheme, we estimate the marginal likelihood by the one-block estimate given in Section 2.1. It should be noted that this scheme provides an entirely different route to finding the marginal likelihood in relation to the one presented in the preceding paragraph. We think it is interesting to see what estimates emerge from these two different approaches to the same problem.

In Figure 2, we report the inefficiency factors under the two schemes for the parameters D_{11}, D_{12}, D_{22}, and σ^2. Observe that the single block M–H scheme lowers the inefficiency factors by a factor of about two. Next, in Table 2 we report the estimates of the marginal likelihood under the two sampling schemes. The first important point to note is that the two sets of estimates are virtually identical, thus showing that the direct Chib method and its extension to M–H chains produce the same answer, when both can be applied, and in parallel with the evidence presented previously, that the scheme that produces the lower inefficiency factors (here the single block scheme) produces the marginal likelihood estimate with the smaller numerical standard error. We should note, however, that not all of the difference in the numerical standard errors is due to the effect of blocking because some of it comes from the difference in the size of the posterior ordinates that are estimated under the two schemes: 4-dimensional in the one-block scheme versus 10-dimensional in the multiple-block scheme.

On the basis of these experiments, which cover important model situations, one may conclude that the marginal likelihood estimation method in this article appears to be robust to changes in the simulation sample sizes, blocking schemes, and sampling methods. For this method, the numerical standard error of the marginal likelihood estimate is lower for the schemes that produce the lower inefficiency factors.

3.3 Poisson Longitudinal Regression

We now consider the calculation of the marginal likelihood in a nonlinear latent variable problem where the likelihood calculation is quite complicated and latent variables cannot be avoided in the MCMC simulation. With this example, which has $n = 58$ blocks of latent variables that must each be sam-

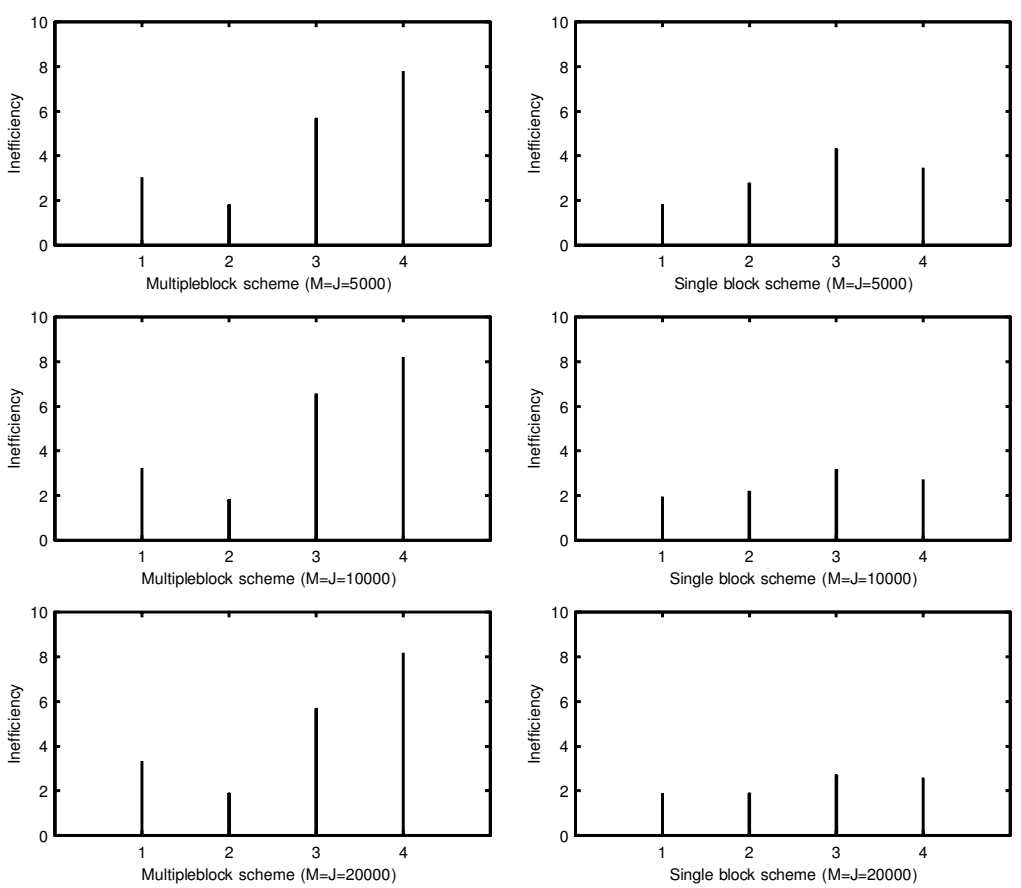

Figure 2. Gaussian Longitudinal Model. Inefficiency factors under two different sampling schemes and choices of (M,J) for the parameters D_{11}, D_{12}, D_{22}, and σ^2.

Table 2. Effect of Blocking Scheme in the Longitudinal Model on Log-marginal Likelihood Estimates

	Type of blocking	
(M, J)	Multiple blocks	Single block
(5,000, 5,000)	−3,577.551 (.028)	−3,577.578 (.009)
(10,000, 10,000)	−3,577.565 (.020)	−3,577.566 (.006)
(20,000, 20,000)	−3,577.575 (.014)	−3,577.574 (.006)

pled by a M–H step, we are able to illustrate the point made in Section 2 that the intractability of the latent variable distributions does not alter the scheme for estimating the marginal likelihood.

The model of interest, taken from Diggle, Liang, and Zeger (1995), is concerned with the modeling of seizure counts $\{y_{it}\}$ for each of $i = 1, \ldots, 59$ epileptics measured first over an 8-week baseline period ($t = 0$) and then over four subsequent 2-week periods $t = 1, \ldots, 4$. At the end of the baseline period, each patient is randomly assigned to either receive the drug progabide or a placebo. After removing observation 49, we fit the model:

$$\begin{aligned} y_{it}|\boldsymbol{\beta}, \mathbf{b}_i &\sim \text{Poisson}(\lambda_{it}) \\ \ln(\lambda_{it}) &= \ln \tau_{it} + \beta_1 x_{it1} + \beta_2 x_{it1} x_{it2} + b_{i1} + b_{i2} x_{it2} \\ \mathbf{b}_i &\sim \mathcal{N}_2(\boldsymbol{\eta}, \mathbf{D}), \end{aligned}$$

where the covariate x_{it1} is an indicator for treatment status, x_{it2} is an indicator of period (0 if baseline and 1 otherwise), τ_{it} is the offset that is equal to 8 in the baseline period and 2 otherwise, and $\mathbf{b}_i = (b_{i1}, b_{i2})$ are latent random effects. This specification of the model produces the likelihood function

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{D}) &= \prod_{i=1}^{58} \int \phi_2(\mathbf{b}_i|\boldsymbol{\eta}, \mathbf{D}) \prod_{t=0}^{4} p(y_{it}|\boldsymbol{\beta}, \mathbf{b}_i)\, d\mathbf{b}_i \\ &\equiv \prod_{i=1}^{58} \int \phi_2(\mathbf{b}_i|\boldsymbol{\eta}, \mathbf{D}) f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)\, d\mathbf{b}_i, \end{aligned} \tag{19}$$

where $p(\cdot)$ is the Poisson mass function with mean λ_{it}. We now follow Chib, Greenberg, and Winkelmann (1998) and conduct the posterior MCMC simulations with the full conditional distributions of $\boldsymbol{\beta}$, $\{\mathbf{b}_i\}$, $\boldsymbol{\eta}$, and $\mathbf{D}^{-1}$ where the $\boldsymbol{\beta}$ and $\{\mathbf{b}_i\}$ blocks are each updated by M–H steps. If we let $\boldsymbol{\beta} \sim \mathcal{N}_2(\mathbf{0}, 100 I_2)$, $\boldsymbol{\eta} \sim \mathcal{N}_2(\mathbf{0}, 100 I_2)$, and $\mathbf{D}^{-1} \sim \mathcal{W}_2(4, I_2)$, then the complete MCMC algorithm for simulating the posterior distribution is defined as follows.

Longitudinal Poisson Model: MCMC Algorithm

1. Calculate the parameters $(\mathbf{m}_0, \mathbf{V}_0)$ as the mode and inverse of the negative Hessian of the mode of $\log \phi_2(\boldsymbol{\beta}|\mathbf{0}, 100 I_2) + \sum_{i=1}^{58} \log f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)$, propose $\boldsymbol{\beta}' \sim f_T(\boldsymbol{\beta}|\mathbf{m}_0, \mathbf{V}_0, \nu) \equiv q(\boldsymbol{\beta}|\mathbf{y}, \{\mathbf{b}_i\})$ and move to $\boldsymbol{\beta}'$ with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}'|\mathbf{y}, \{\mathbf{b}_i\}) = \min\left\{ \frac{\prod_{i=1}^{58} f(\mathbf{y}_i|\boldsymbol{\beta}', \mathbf{b}_i) \phi_2(\boldsymbol{\beta}'|\mathbf{0}, 100 I_2)}{\prod_{i=1}^{58} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) \phi_2(\boldsymbol{\beta}|\mathbf{0}, 100 I_2)} \times \frac{f_T(\boldsymbol{\beta}|\mathbf{m}_0, \mathbf{V}_0, 15)}{f_T(\boldsymbol{\beta}'|\mathbf{m}_0, \mathbf{V}_0, 15)}, 1 \right\}.$$

2. Calculate the parameters $(\mathbf{m}_i, \mathbf{V}_i)$ as the mode and inverse of the negative Hessian of the mode of $\log \phi_2(\mathbf{b}_i|\boldsymbol{\eta}, \mathbf{D}) + \log f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)$, propose $\mathbf{b}_i' \sim f_T(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)$ and move to $\mathbf{b}_i'$ with probability

$$\alpha_i = \min\left\{ \frac{f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i') \phi_2(\mathbf{b}_i'|\boldsymbol{\eta}, \mathbf{D})}{f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) \phi_2(\mathbf{b}_i|\boldsymbol{\eta}, \mathbf{D})} \frac{f_T(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)}{f_T(\mathbf{b}_i'|\mathbf{m}_i, \mathbf{V}_i, \nu)}, 1 \right\}.$$

3. Sample $\boldsymbol{\eta} \sim \mathcal{N}_2(\boldsymbol{\eta}|\hat{\boldsymbol{\eta}}, M)$ where $\mathbf{M} = (100^{-1} I_2 + 58 \mathbf{D}^{-1})^{-1}$ and $\hat{\boldsymbol{\eta}} = \mathbf{M} \sum_{i=1}^{58} \mathbf{D}^{-1} \mathbf{b}_i$.

4. Sample

$$\mathbf{D}^{-1} \sim \mathcal{W}_2\left(62, [I + \sum_{i=1}^{58} (\mathbf{b}_i - \boldsymbol{\eta})(\mathbf{b}_i - \boldsymbol{\eta})']^{-1}\right).$$

To estimate the marginal likelihood, we observe that this MCMC scheme is a special case of the setup in Section 2.3 with $B = 3$, $\boldsymbol{\theta}_1 = \mathbf{D}^{-1}$, $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$, and $\boldsymbol{\theta}_3 = \boldsymbol{\eta}$, where the additional blocks of latent variables can be integrated out in the manner described in Section 2.2. We begin by decomposing the posterior ordinate as

$$\pi(\mathbf{D}^{-1*}, \boldsymbol{\beta}^*, \boldsymbol{\eta}^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\, \pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*)\, \pi(\boldsymbol{\eta}^*|\mathbf{y}, \mathbf{D}^*, \boldsymbol{\beta}^*),$$

where only the 2-dimensional reduced ordinate $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*)$ cannot be estimated by the Rao–Blackwell device. Specifically, an estimate of the first ordinate can be found by averaging the Wishart density over draws on $\{\mathbf{b}_i\}$ and $\boldsymbol{\eta}$ from the full run. Then, from Section 2.3, our estimate of the second ordinate is given by

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*) = \frac{M^{-1} \sum_{g=1}^{M} \alpha(\boldsymbol{\beta}^{(g)}, \boldsymbol{\beta}^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\}) q(\boldsymbol{\beta}^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\})}{J^{-1} \sum_{j=1}^{J} \alpha(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(j)}|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})} \tag{20}$$

where the draws in the numerator are from a reduced run with the full conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\{\mathbf{b}_i\}$, conditioned on $\mathbf{D}^*$. The draws in the denominator are from a second reduced run with the full conditional distributions of $\{\mathbf{b}_i\}$ and $\boldsymbol{\eta}$, conditioned on $(\mathbf{D}^*, \boldsymbol{\beta}^*)$ with an appended step in which $\boldsymbol{\beta}^{(j)}$ is drawn from $q(\boldsymbol{\beta}|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})$. The draws on $\{\mathbf{b}_i\}$ in the latter run are also used to average the normal density of $\boldsymbol{\eta}$ to produce an estimate of the third ordinate. The log-marginal likelihood is then obtained from (18) after the likelihood ordinate $f(\mathbf{y}|\boldsymbol{\beta}^*, \boldsymbol{\eta}^*, \mathbf{D}^*)$ is found by estimating the integral in (19) by an importance sampling method.

We compare the performance of the new approach by also computing the marginal likelihood as in Chib, Greenberg, and Winkelmann (1998) where a kernel density estimation method is used to calculate $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*)$. Because the dimension of $\boldsymbol{\beta}$

is small, we expect that the two methods should agree closely. In fact, the new method produces an estimate of -915.23 based on $M = J = 10{,}000$ and the older method gives the value of -915.49, with numerical standard errors from both methods of approximately .1. As the dimension of $\boldsymbol{\beta}$ increases, the kernel density estimate of the ordinate $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*)$ will become less accurate and one will have to rely on the approach developed in this article.

3.4 Multivariate Probit Model

The final example again involves a large number of latent variables in the nonlinear model setting of a model for correlated binary data. The model of interest is the *multivariate probit model,* which we fit using data from the Panel Study of Income Dynamics of the University of Michigan. The response variable is the labor force participation decision of 520 married women, in the age range 25–62, each observed over the 7-year span 1976–1982. Under the fitted model, the marginal probability of participation status of the ith woman at the jth time point is given by

$$\Pr(y_{ij} = 1|\boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}), \qquad i \le 520, j \le 7,$$

where x_1 is the ith subject's education measured as the number of grades completed, x_2 is total family income excluding the woman's earnings, in thousands of dollars, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and Φ is the distribution function of the standard normal distribution. The joint probability of a particular vector $\mathbf{y}_i : 7 \times 1$ representing the sequence of participation status indicators for the ith subject, conditioned on the parameters $\boldsymbol{\beta}$ and a correlation matrix $\boldsymbol{\Sigma}$ is

$$\Pr(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{B_{i7}} \cdots \int_{B_{i1}} \phi_J(\mathbf{z}_i|\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})\, d\mathbf{z}_i, \qquad (21)$$

where B_{ij} is the interval $(0, \infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{i7})$ is a vector of latent variables with distribution

$$\mathbf{z}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_7(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

$\mathbf{X}_i = (x_{i0}, \ldots, x_{i2})'$ is a 7×3 matrix of the covariates augmented with a vector of 1s and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_2)'$ are the regression parameters. The 21 free elements of the correlation matrix $\boldsymbol{\Sigma}$ are denoted by $\boldsymbol{\sigma} = (\sigma_{21}, \sigma_{31}, \sigma_{32}, \sigma_{41}, \sigma_{42}, \sigma_{43}, \ldots, \sigma_{71}, \ldots, \sigma_{76})$. The likelihood of this model is difficult to compute because the integral in (21), which defines the likelihood contribution of the ith subject, is not available in closed form. This necessitates the use of latent data in the MCMC sampling.

We now consider how well our method performs in this context. Under the prior distributions $\pi(\boldsymbol{\beta}) = \phi_3(\boldsymbol{\beta}|\mathbf{0}, 10I_4)$ and $\pi(\boldsymbol{\sigma}) \propto \phi_{21}(\boldsymbol{\sigma}|.3i_{21}, .2I_{21})I_S$, where i_{21} is a vector of ones and S is the subset of R^{21} that leads to a positive definite correlation matrix, the posterior distribution is sampled using the algorithm of Chib and Greenberg (1998), which is based on the method of Albert and Chib (1993), by the $(520 \times 7) + 2$ conditional distributions

$$\{\pi(\{z_{ij}\}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma})\} \qquad (i = 1, \ldots, 520,\ j = 1, \ldots, 7);$$
$$\pi(\boldsymbol{\beta}|\mathbf{y}, \{\mathbf{z}_i\}, \boldsymbol{\Sigma}); \qquad \pi(\boldsymbol{\sigma}|\mathbf{y}, \{\mathbf{z}_i\}, \boldsymbol{\beta}),$$

where the last distribution, which is high-dimensional, is sampled in one block by a conditional tailored M–H step.

Multivariate probit model: MCMC algorithm

1. Sample for $i \le 520,\ j \le 7$

$$z_{ij} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{ij}, v_{ij}) & \text{if } y_{ij} = 1 \\ \mathcal{TN}_{(-\infty,0)}(\mu_{ij}, v_{ij}) & \text{if } y_{ij} = 0 \end{cases}$$

where $\mu_{ij} = \mathrm{E}(z_{ij}|z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $v_{ij} = \mathrm{var}(z_{ij}|z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, and $\mathcal{TN}_{(a,b)}^{(\mu,s)}$ denotes the normal distribution with parameters (μ, s) truncated to the interval (a, b).

2. Sample $\boldsymbol{\beta} \sim \mathcal{N}_k(\hat{\boldsymbol{\beta}}, \mathbf{B}_n)$, where $\hat{\boldsymbol{\beta}} = \mathbf{B}_n(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{520} \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}z_i)$ and $\mathbf{B}_n = (10^{-1}I_3 + \sum_{i=1}^{520} \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i)^{-1}$ are the usual updates based on the complete data.

3. Calculate $\mathbf{m} = \arg\max_{\boldsymbol{\sigma}} \log \phi_{21}(\boldsymbol{\sigma}|.3i_{21}, .2I_{21}) + \sum_{i=1}^{520} \log \phi_{21}(\mathbf{z}_i|\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\mathbf{V}$ the negative inverse of the Hessian of the log-target at the mode, propose $\boldsymbol{\sigma}' \sim f_T(\boldsymbol{\sigma}|\mathbf{m}, \mathbf{V}, \nu)$ and move to $\boldsymbol{\sigma}'$ with probability

$$\alpha = \min\left\{ \frac{\phi_{21}(\boldsymbol{\sigma}'|.3i_{21}, .2I_{21}) \prod_{i=1}^{520} \phi_{21}(\mathbf{z}_i|\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}')I[\boldsymbol{\sigma}' \in S]}{\phi_{21}(\boldsymbol{\sigma}|.3i_{21}, .2I_{21}) \prod_{i=1}^{520} \phi_{21}(\mathbf{z}_i|\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})} \times \frac{f_T(\boldsymbol{\sigma}|\mathbf{m}, \mathbf{V}, \nu)}{f_T(\boldsymbol{\sigma}'|\mathbf{m}, \mathbf{V}, \nu)}, 1 \right\}$$

It should be clear that this situation is similar to the one discussed in Section 2.2 except that the latent variables are sampled not in one block but in several blocks. If we decompose the posterior ordinate at the posterior mean as

$$\pi(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^*|\mathbf{y}) = \pi(\boldsymbol{\sigma}^*|\mathbf{y})\pi(\boldsymbol{\beta}^*|\mathbf{y}, \boldsymbol{\sigma}^*),$$

then the first ordinate is available from (14) and the second ordinate as the average

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{y}, \boldsymbol{\Sigma}^*) = M^{-1} \sum_{g=1}^{M} \phi_3(\boldsymbol{\beta}^*|\hat{\boldsymbol{\beta}}^{*(g)}, \mathbf{B}_n^*)$$

where $\hat{\boldsymbol{\beta}}^{*(g)} = \mathbf{B}_n^*(\sum_{i=1}^{520} \mathbf{X}_i'\boldsymbol{\Sigma}^{*-1}\mathbf{z}_i^{(g)})$, $\mathbf{B}_n^* = (10^{-1}I_3 + \sum_{i=1}^{520} \mathbf{X}_i'\boldsymbol{\Sigma}^{*-1}\mathbf{X}_i)^{-1}$ and $\mathbf{z}_i^{(g)}$ are draws from a reduced MCMC run with $\boldsymbol{\Sigma}$ fixed at $\boldsymbol{\Sigma}^*$. We are particularly interested in seeing how the proposed method is able to estimate the marginal likelihood if the 21-dimensional ordinate $\pi(\boldsymbol{\sigma}^*|\mathbf{y})$ is estimated in one block. For comparison purposes we also utilize the approach of Chib and Greenberg (1998), where the posterior ordinate is estimated from

$$\pi(\boldsymbol{\sigma}_1^*|\mathbf{y})\pi(\boldsymbol{\sigma}_2^*|\mathbf{y}, \boldsymbol{\sigma}_1^*)\pi(\boldsymbol{\sigma}_3^*|\mathbf{y}, \boldsymbol{\sigma}_1^*, \boldsymbol{\sigma}_2^*)\pi(\boldsymbol{\sigma}_4^*|\mathbf{y}, \boldsymbol{\sigma}_1^*, \boldsymbol{\sigma}_2^*, \boldsymbol{\sigma}_3^*),$$

where $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$, and $\boldsymbol{\sigma}_3$ are 6-dimensional subblocks of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}_4$ is 3-dimensional. Each of the conditional ordinates in the latter expression is estimated by kernel smoothing applied to 20,000 values of $\boldsymbol{\sigma}_i$ generated from $\pi(\boldsymbol{\sigma}_i|\mathbf{y}, \boldsymbol{\sigma}_1^*, \ldots, \boldsymbol{\sigma}_{i-1}^*)$ in a reduced Markov chain Monte Carlo run. In relation to the one $\boldsymbol{\sigma}$ block setup, the latter sampler requires three additional reduced runs and consumes about twice as much time. Interestingly, however, the two setups produce virtually the same answer. From the one $\boldsymbol{\sigma}$ block method based on $M = J = 20{,}000$, the marginal likelihood is found to be $-1{,}550.533$ with a numerical standard error of .842. Then, the

multiple $\boldsymbol{\sigma}$ block setup produces the value $-1{,}550.45$ with a numerical standard error of .278. The larger numerical standard error of the one $\boldsymbol{\sigma}$ block method is a consequence of the fact that the sampling of $\boldsymbol{\sigma}$ from $\pi(\boldsymbol{\sigma}|\mathbf{y}, \{\mathbf{z}_i\}, \boldsymbol{\beta})$ produces higher inefficiency factors for each component of $\boldsymbol{\sigma}$ in relation to the output on $\boldsymbol{\sigma}$ when it is generated in three 6-dimensional blocks and one 3-dimensional block, which shows once again that the more efficient MCMC simulation routine reduces the numerical standard error of the marginal likelihood estimate.

3.5 Additional Examples and Alternative Methods

The method developed in this article has been applied to other important models. For example, we have conducted an analysis of the method in the context of ordinal probit models, autoregressive and moving average (ARMA) models, and parametric models for survival data such as the Weibull and the log-logistic. We find support for the same general conclusions presented previously. For instance, in our ordinal data example, a 16-dimensional posterior ordinate is estimated two ways: from the output of a single block M–H algorithm and from the output of a two-block algorithm. In this case, the single block algorithm produces higher inefficiency factors and, in keeping with the results presented previously, a higher numerical standard error in the marginal likelihood estimation.

In addition, we have considered some of the alternative methods that are discussed in the introduction. One such alternative is the method of kernel smoothing to estimate the posterior ordinate. For high-dimensional problems, we find that the kernel-based estimate exhibits two deficiencies: larger numerical standard errors and significant bias that tends to dissipate slowly as a function of the simulation sample size. For example, in the logit model of Section 3.1, the kernel based estimate is -141.722 after 5,000 draws, and -142.229 after 40,000 draws; and, standard errors are approximately 10 times larger than those produced by the proposed method. The kernel estimates can be made more accurate by breaking up the parameter vector into smaller blocks, as is done in the multivariate probit example, and then proceeding with additional reduced runs to produce the correct sample draws. This procedure reduces the bias and lowers the numerical standard error of the marginal likelihood estimate but takes increasingly more time as the size of each MCMC block is increased. Thus to apply this method one would usually need more blocks in the MCMC sampling than would arise from the natural full conditional structure of the model.

Furthermore, we have examined the marginal likelihood estimation methods proposed by Gelfand and Dey (1994) and Meng and Wong (1996) in the context of the problems that are discussed in this article. Although these methods both require the use of certain auxiliary functions, each produces results in close agreement with those in Tables 1 and 2. This is what one might expect given that both models are relatively simple and the sample sizes are relatively large. In the last two models, however, these methods are computationally much more expensive than our method because of the required evaluation of the likelihood for each sampled draw. In the multivariate probit model discussed previously, for example, a single likelihood evaluation, which is all that is required in our approach, takes about 2 minutes of computing time on a 550 MHz machine. Multiple evaluations of the likelihood are therefore extremely costly. It is not possible to circumvent this burden by keeping the approximately 3,600 latent variables in the last example because the consequent increase in the dimension of the ordinates leads to inefficient estimates. Model space methods, on the other hand, have a different underpinning and provide a complementary set of tools for finding Bayes factors. In general, these methods cannot be competitive with a direct marginal likelihood estimation method when the number of models being compared is small because it is easier to fit each model directly than to design a new mega-simulation procedure with its attendant tuning and other costs. Han and Carlin (2000), in the context of our second example, make the same point and report favorably on our method in relation to the methods of Green (1995) and Carlin and Chib (1995). Nonetheless, a comparative study of these methods in more general settings, such as those in our last two examples, will be desirable once suitable model space methods for those models have been worked out.

4. CONCLUDING REMARKS

In this article, we have derived and illustrated an extended version of the Chib method of computing model marginal likelihoods for Bayesian model comparisons based on the output of Metropolis–Hastings MCMC chains. By completing the Chib method we now have a framework for calculating marginal likelihoods in practically all models that are fit by Markov chain Monte Carlo methods. One virtue of the approach is that it is based on the programming that is done to simulate the posterior distribution. Thus once the basic programming has been done, one can find the marginal likelihood of a given collection of models by rearrangement of existing code and, importantly, without further tuning of the MCMC algorithm. In experiments involving the logit model for binary data, hierarchical random effects model for clustered Gaussian data, Poisson regression model for clustered count data, and the multivariate probit model for correlated binary data, we have illustrated the performance and implementation of the method. In our examples, the method is robust to changes in the blocking schemes and proposal densities that are used to sample the posterior distribution. Furthermore, the sampling scheme that is efficient for sampling the posterior distribution, as measured by the inefficiency factors, is also efficient for estimating the marginal likelihood.

[Received July 1999. Revised June 2000.]

REFERENCES

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

Brown, B. W. (1980), "Prediction Analyses for Binary Data," in *Biostatistics Casebook*, eds. R. J. Miller, B. Efron, B. W. Brown, and L. E. Moses, New York: Wiley.

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

Chen, M.-H., and Shao, Q.-M. (1997), "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants," *The Annals of Statistics*, 25, 1563–1594.

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

Chib, S., and Carlin, B. P. (1999), "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hastings Algorithm," *American Statistician*, 49, 327–335.

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 2, 347–361.

Chib, S., Greenberg, E., and Winkelmann, R. (1998), "Posterior Simulation and Bayes Factors in Panel Count Data Models," *Journal of Econometrics*, 86, 33–54.

Chib, S., Nardari, F., and Shephard, N. (1999), "Analysis of High Dimensional Multivariate Stochastic Volatility Models," *Technical report*, Olin School of Business, Washington University, St. Louis, MO.

DiCiccio, T. J., Kass, R. E., Raftery, A. E., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Diggle, P., Liang, K.-K., and Zeger, S. L. (1995), *Analysis of Longitudinal Data*, Oxford: 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.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Han, C., and Carlin, B. (2000), "MCMC methods for Computing Bayes Factors: A Comparative Review," *Research report* 2000–001, Division of Biostatistics, University of Minnesota.

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

Meng, X.-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Newey, W. K., and West, K. D. (1987), "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Ritter, C., and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy–Gibbs sampler," *Journal of the American Statistical Association*, 87, 861–868.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association,* 82, 528–549.

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

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.