# Marginal Likelihood From the Gibbs Output

Siddhartha CHIB

In the context of Bayes estimation via Gibbs sampling, with or without data augmentation, a simple approach is developed for computing the marginal density of the sample data (marginal likelihood) given parameter draws from the posterior distribution. Consequently, Bayes factors for model comparisons can be routinely computed as a by-product of the simulation. Hitherto, this calculation has proved extremely challenging. Our approach exploits the fact that the marginal density can be expressed as the prior times the likelihood function over the posterior density. This simple identity holds for any parameter value. An estimate of the posterior density is shown to be available if all complete conditional densities used in the Gibbs sampler have closed-form expressions. To improve accuracy, the posterior density is estimated at a high density point, and the numerical standard error of resulting estimate is derived. The ideas are applied to probit regression and finite mixture models.

KEY WORDS: Bayes factor; Estimation of normalizing constant; Finite mixture models; Linear regression; Markov chain Monte Carlo; Markov mixture model; Multivariate density estimation; Numerical standard error; Probit regression; Reduced conditional density.

## 1. INTRODUCTION

The advent of Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith 1990, Tanner and Wong 1987) to simulate posterior distributions has virtually revolutionized the practice of Bayesian statistics. For the most part, these methods have been used for estimation and out-of-sample prediction, because both of those problems are easily solved given a sample of draws from the posterior distribution. On the other hand, the problem of calculating the marginal likelihood, which is the normalizing constant of the posterior density and an input to the computation of Bayes factors (see, for example, Berger 1985, Kass and Raftery 1995, or O'Hagan 1994), has proved extremely challenging. This is because the marginal likelihood is obtained by integrating the likelihood function with respect to the prior density, whereas the MCMC method produces draws from the posterior.

One way to deal with this problem is to compute Bayes factors without attempting to calculate the marginal likelihood by introducing a model indicator into the list of unknown parameters. Work along these lines has been reported by Carlin and Polson (1991), Carlin and Chib (1995), and many others. To use these methods, however, it is necessary to specify all of the competing models at the outset, which may not be always possible, and to carefully specify certain tuning constants to ensure that the simulation algorithm mixes suitably in model space. In this article, therefore, we concern ourselves with methods that directly address the calculation of the marginal likelihood. Suppose that $f(\mathbf{y}|\boldsymbol{\theta}_k, M_k)$ is the density function of the data $\mathbf{y} = (y_1, \ldots, y_n)$ under model $M_k$ $(k = 1, 2, \ldots, K)$ given the model-specific parameter vector $\boldsymbol{\theta}_k$. Let the prior density of $\boldsymbol{\theta}_k$ (assumed to be proper) be given by $\pi(\boldsymbol{\theta}_k|M_k)$, and let $\{\boldsymbol{\theta}_k^{(g)}\} \equiv \{\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(G)}\}$ be $G$ draws from the posterior density $\pi(\boldsymbol{\theta}_k|\mathbf{y}, M_k)$ obtained using a MCMC method, say the Gibbs sampler. Newton and Raftery (1994) showed that the marginal likelihood (equivalently, the marginal density of $\mathbf{y}$) under model $M_k$, that is,

$$m(\mathbf{y}|M_k) = \int f(\mathbf{y}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)\, d\boldsymbol{\theta}_k, \tag{1}$$

can be estimated as

$$\hat{m}_{\mathrm{NR}} = \left\{ \frac{1}{G} \sum_{g=1}^{G} \left( \frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_k^{(g)}, M_k)} \right) \right\}^{-1}, \tag{2}$$

which is the harmonic mean of the likelihood values. Although this estimate is a simulation-consistent estimate of $m(\mathbf{y}|M_k)$, it is not stable, because the inverse likelihood does not have finite variance. But consider the quantity proposed by Gelfand and Dey (1993):

$$\hat{m}_{\mathrm{GD}} = \left\{ \frac{1}{G} \sum_{g=1}^{G} \left( \frac{p(\boldsymbol{\theta}_k^{(g)}}{f(\mathbf{y}|\boldsymbol{\theta}_k^{(g)}, M_k)\pi(\boldsymbol{\theta}_k^{(g)}|M_k)} \right) \right\}^{-1}, \tag{3}$$

where $p(\boldsymbol{\theta})$ is a density with tails thinner than the product of the prior and the likelihood. This can be shown to have the property that $\hat{m}_{\mathrm{GD}} \to m(\mathbf{y}|M_k)$ as $G$ becomes large without the instability of $\hat{m}_{\mathrm{NR}}$. Nonetheless, this approach requires a tuning function, which can be quite difficult to determine in high-dimensional problems, and subsequent monitoring to ensure that the numbers are stable. In fact, we have found that the somewhat obvious choices of $p(\cdot)$—a normal density or $t$ density with mean and covariance equal to the posterior mean and covariance—do not necessarily satisfy the thinness requirement. Other attempts to modify the harmonic mean estimator, though requiring samples from both the prior and posterior distributions, have been discussed by Newton and Raftery (1994).

The purpose of this article is to demonstrate that a simple approach to computing the marginal likelihood and the Bayes factor is available that is free of the problems just described. This approach is developed in the setting where the

Siddhartha Chib is Professor of Econometrics, John M. Olin School of Business, Washington University, St. Louis, MO 63130. This article has benefited from valuable comments of two anonymous referees, the associate editor, and the editor. In addition, discussions with Jim Albert, Ed Greenberg, and Radford Neal are gratefully acknowledged.

© 1995 American Statistical Association
Journal of the American Statistical Association
December 1995, Vol. 90, No. 432, Theory and Methods

Gibbs sampling algorithm, with or without data augmentation, has been used to provide a sample of draws from the posterior distribution. To compute the marginal density by our approach, it is necessary that all integrating constants of the full conditional distributions in the Gibbs sampler be known. This requirement is usually satisfied in models fit with conjugate priors and covers almost all applications of the Gibbs sampler that have appeared in the literature.

The rest of the article is organized as follows. Section 2 presents the approach, and Section 3 illustrates the derivation of the numerical standard error of the estimate. Section 4 presents applications of the approach, first for variable selection in probit regression and then for model comparisons in finite mixture models. The final section contains brief concluding remarks.

## 2. THE APPROACH

Suppress the model index $k$ and consider the situation wherein $f(\mathbf{y}|\boldsymbol{\theta})$ is the sampling density (likelihood function) for the given model and $\pi(\boldsymbol{\theta})$ is the prior density. To allow for the possibility that posterior simulation requires data augmentation, let $\mathbf{z}$ denote latent data and suppose that for a given set of vector blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B)$, the Gibbs sampling algorithm is applied to the set of $(B + 1)$ complete conditional densities,

$$\{\pi(\boldsymbol{\theta}_r|\mathbf{y}, \boldsymbol{\theta}_s(s \neq r), \mathbf{z})\}_{r=1}^{B}, \qquad p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}). \tag{4}$$

The objective is to compute the marginal density $m(\mathbf{y}|M_k)$ from the output $\{\boldsymbol{\theta}^{(g)}, \mathbf{z}^{(g)}\}_{g=1}^{G}$ obtained from (4).

The approach developed here consists of two related ideas. First, $m(\mathbf{y})$, by virtue of being the normalizing constant of the posterior density, can be written as

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{y})}, \tag{5}$$

where the numerator is just the product of the sampling density and the prior, with all integrating constants included, and the denominator is the posterior density of $\boldsymbol{\theta}$. It is worthwhile to refer to this simple identity, which holds for any $\boldsymbol{\theta}$, as the *basic marginal likelihood identity* (BMI). Second, for a given $\boldsymbol{\theta}$ (say $\boldsymbol{\theta}^*$), the posterior ordinate $\pi(\boldsymbol{\theta}^*|\mathbf{y})$ can be estimated by exploiting the information in the collection of complete conditional densities $\{\pi(\boldsymbol{\theta}_r|\mathbf{y}, \boldsymbol{\theta}_s(s \neq r), \mathbf{z})\}_{r=1}^{B}$. The technique for doing so is described later, but for the present, if the posterior density estimate at $\boldsymbol{\theta}^*$ is denoted by $\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y})$, then the proposed estimate of the marginal density, on the computationally convenient logarithm scale, is

$$\ln \hat{m}(\mathbf{y}) = \ln f(\mathbf{y}|\boldsymbol{\theta}^*) + \ln \pi(\boldsymbol{\theta}^*) - \ln \hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}). \tag{6}$$

It is important to observe the simplicity and benefits of this expression: all it requires is the evaluation of the log-likelihood function and the prior and an estimate of posterior ordinate. The estimate does not suffer from any instability problem, because it is a density value that is averaged rather than its inverse. In addition, the entire estimation (simulation) error arises from the estimation of the posterior ordinate, and this simulation error can be derived, as shown in Section 3. It is now time to examine the method for calculating the posterior density estimate from the Gibbs output.

### 2.1 Estimation of $\pi(\boldsymbol{\theta}^*|\mathbf{y})$.

Consider now the estimation of the multivariate density $\pi(\boldsymbol{\theta}^*|\mathbf{y})$ and the selection of the point $\boldsymbol{\theta}^*$. As was pointed out, the BMI expression holds for any $\boldsymbol{\theta}$, and thus the choice of the point is not critical, but efficiency considerations dictate that for a given number of posterior draws, the density is likely to be more accurately estimated at a high density point, where more samples are available, than at a point in the tails. It should be noted that a modal value such as the posterior mode, or the maximum likelihood estimate, can be computed from the Gibbs output, at least approximately, if it is easy to evaluate the log-likelihood function for each draw in the simulation. Alternatively, one can make use of the posterior mean provided that there is no concern that it is a low density point.

We now explain how the posterior density ordinate can be estimated from the Gibbs output, starting with a canonical situation consisting of two blocks of parameters before turning to the general case. We show that the proposed multivariate density estimation method is easy to implement, requires only the available complete conditional densities, and produces a simulation consistent estimate of the posterior ordinate.

*2.1.1 Two Vector Blocks.* Suppose that Gibbs sampling is applied to the complete conditional densities

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}); \qquad p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}),$$

which is the setting of Tanner and Wong (1987). Let the output from the Gibbs algorithm be given by $\{\boldsymbol{\theta}^{(g)}, \mathbf{z}^{(g)}\}_{g=1}^{G}$ and suppose that $\boldsymbol{\theta}^*$ is the selected point. If the posterior density is written as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) p(\mathbf{z}|\mathbf{y})\, d\mathbf{z},$$

then it follows that an appropriate Monte Carlo estimate of $\pi(\boldsymbol{\theta}|\mathbf{y})$ at $\boldsymbol{\theta}^*$ is

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) = G^{-1} \sum_{g=1}^{G} \pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{z}^{(g)}), \tag{7}$$

because $\mathbf{z}^{(g)}$ is a draw from the distribution $\mathbf{z}|\mathbf{y}$. Gelfand and Smith (1990) referred to this technique as Rao–Blackwellization and argued that it improves on the multivariate kernel method (Scott 1992). Also, under regularity conditions, the estimate is simulation consistent; that is, $\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) \to \pi(\boldsymbol{\theta}^*|\mathbf{y})$ as $G$ becomes large, almost surely, as a consequence of the ergodic theorem (Tierney 1994). Substituting the estimate of the posterior ordinate into (6) gives

the following estimate of the marginal likelihood:

$$\ln \hat{m}(\mathbf{y}) = \ln f(\mathbf{y}|\boldsymbol{\theta}^*) + \ln \pi(\boldsymbol{\theta}^*) - \ln \left\{ G^{-1} \sum_{g=1}^{G} \pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{z}^{(g)}) \right\}.$$

This simple expression can be used for a large class of models, including the probit regression model discussed later. Observe that the calculation amounts to evaluating the likelihood, the prior, and the "complete data" posterior density at the point $\boldsymbol{\theta}^*$.

*2.1.2 Three Vector Blocks.* An even larger class of models can be covered by slightly generalizing the Tanner and Wong structure. Suppose that the Gibbs sampler is defined through the complete conditional densities

$$\pi(\boldsymbol{\theta}_1|\mathbf{y}, \theta_2, \mathbf{z}); \qquad \pi(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1, \mathbf{z}); \qquad p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}).$$

Models such as linear regression, linear regression with independent Student-$t$ errors, Zellner's seemingly unrelated regression, and censored regression either fall in this category or are a special case of this structure if $\mathbf{z}$ is absent. Once again, the objective is to estimate $\pi(\boldsymbol{\theta}^*|\mathbf{y})$, which now is expressed as

$$\pi(\boldsymbol{\theta}_1^*|\mathbf{y})\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*), \tag{8}$$

where

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

and

$$\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) = \int \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z})p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)\, d\mathbf{z} \tag{9}$$

is the reduced conditional density ordinate. It should be clear that the normalizing constants of $\pi(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \mathbf{z})$ and $\pi(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1, \mathbf{z})$ must be included in the integration for the decomposition in (8) to be valid. The first ordinate, $\pi(\boldsymbol{\theta}_1^*|\mathbf{y})$, can be estimated in an obvious way, by taking the ergodic average of the full conditional density with the posterior draws of $(\boldsymbol{\theta}_2, \mathbf{z})$, leading to the estimate

$$\hat{\pi}(\boldsymbol{\theta}_1^*|\mathbf{y}) = G^{-1} \sum_{g=1}^{G} \pi(\boldsymbol{\theta}_1^*|\mathbf{y}, \boldsymbol{\theta}_2^{(g)}, \mathbf{z}^{(g)}).$$

A similar technique, with an important twist, can be invoked to obtain the reduced conditional ordinate in (9). Recognize that the draws of $\mathbf{z}$ from the Gibbs sampler are from the distribution $[\mathbf{z}|\mathbf{y}]$ and not from $[\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*]$. Therefore, the complete conditional density of $\boldsymbol{\theta}_2$ cannot be averaged directly. A simple solution is available to deal with this complication: Continue sampling for an additional $G$ iterations with the complete conditional densities

$$\pi(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z}) \quad \text{and} \quad p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2),$$

where in each of these densities, $\boldsymbol{\theta}_1$ is set equal to $\boldsymbol{\theta}_1^*$. From MCMC theory, it can be verified that the draws $\{\mathbf{z}^{(j)}\}$ from this run follow the density $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*)$, as required. Consequently, $\hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) = G^{-1} \sum \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \mathbf{z}^{(j)})$ is a simulation consistent estimate of (9). Although this procedure leads to an increase in the number of iterations, it is important to stress that it does not require new programming and thus is straightforward to implement. Note that the reduced conditional run is not necessary if $\mathbf{z}$ is absent from the sampling. In this case the reduced conditional density of $\boldsymbol{\theta}_2$ is identical to its complete conditional density, and the density estimate reduces to one used by Zellner and Min (1995) in a different context.

Substituting the two density estimates into (6) yields the estimate

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

*2.1.3 General Case.* Although the technique described thus far will apply to many problems of importance, consider the situation with an arbitrary number of blocks. Even in this case, the posterior density ordinate can be estimated rather easily.

Begin by writing the posterior density at the selected point as

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

where the first term is the marginal ordinate, which can be estimated from the draws of the initial Gibbs run, and the typical term is the reduced conditional ordinate $\pi(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{r-1}^*)$. The latter is given by

$$\int \pi(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{r-1}^*, \boldsymbol{\theta}_l(l > r), \mathbf{z})\, d\pi(\boldsymbol{\theta}_{r+1}, \ldots, \boldsymbol{\theta}_B, \mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{r-1}^*), \tag{10}$$

where $\pi$ is being used to denote density and distribution function interchangeably. To estimate this term, continue the sampling with the complete conditional densities of $\{\boldsymbol{\theta}_r, \boldsymbol{\theta}_{r+1}, \ldots, \boldsymbol{\theta}_B, \mathbf{z}\}$, where in each of these full conditional densities, $\boldsymbol{\theta}_s$ is set equal to $\boldsymbol{\theta}_s^*, (s \leq r-1)$. If the draws from the reduced complete conditional Gibbs run are denoted by $\{\boldsymbol{\theta}_r^{(j)}, \boldsymbol{\theta}_{r+1}^{(j)}, \ldots, \boldsymbol{\theta}_B^{(j)}, \mathbf{z}^{(j)}\}$, then an estimate of (10) is

$$\hat{\pi}(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_s^*(s < r)) = G^{-1} \sum_{j=1}^{G} \pi(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{r-1}^*, \boldsymbol{\theta}_l^{(j)}(l > r), \mathbf{z}^{(j)}), \tag{11}$$

whereas an estimate of the joint density is $\prod_{r=1}^{B} \hat{\pi}(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_s^*(s < r))$. The log of the marginal likelihood is

$$\ln \hat{m}(\mathbf{y}) = \ln f(\mathbf{y}|\boldsymbol{\theta}^*) + \ln \pi(\boldsymbol{\theta}^*) - \sum_{r=1}^{B} \ln \hat{\pi}(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_s^*(s < r)). \tag{12}$$

As an illustration of this procedure, suppose that $B = 3$, a situation that arises in longitudinal random effects models

and many other models. Then $\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)$ is estimated as $G^{-1}\sum \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^{(j)}, \mathbf{z}^{(j)})$, where the draws $\{\boldsymbol{\theta}_3^{(j)}, \mathbf{z}^{(j)}\}$ are obtained by continuing the Gibbs sampler with

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

and

$$\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3).$$

Finally, additional $G$ iterations with the densities

$$\pi(\boldsymbol{\theta}_3|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \mathbf{z}) \quad \text{and} \quad \pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3)$$

produce draws $\{\mathbf{z}^{(j)}\}$ that follow the distribution $[\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*]$. These draws yield an estimate $\pi(\boldsymbol{\theta}_3^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)$. This technique is illustrated in Section 4.2 for mixture models.

### 2.2 Bayes Factor Estimate

To compute the Bayes factor for any two models $k$ and $l$—that is, $m(\mathbf{y}|M_k)/m(\mathbf{y}|M_l)$—the calculation described earlier is repeated for all models, and the following estimate is used:

$$\hat{B}_{kl} = \exp\{\ln \hat{m}(\mathbf{y}|M_k) - \ln \hat{m}(\mathbf{y}|M_l)\}.$$

An estimate of the posterior odds of any two models is given by multiplying the estimated Bayes factor by the prior odds.

### 2.3 Remarks

In some situations there are two sets of latent vectors $(\mathbf{z}, \boldsymbol{\psi})$ such that the density $f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi})\, d\mathbf{z}$ is available in closed form but the likelihood $f(\mathbf{y}|\boldsymbol{\theta}) = \int f(\mathbf{y}, \boldsymbol{\psi}|\boldsymbol{\theta})\, d\boldsymbol{\psi}$ is not. This occurs, for example, in discrete response data models with random effects. To analyze this situation, one can use the BMI expression

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}, \boldsymbol{\psi})}{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}.$$

Both the numerator and denominator can be evaluated at the point $(\boldsymbol{\theta}^*, \boldsymbol{\psi}^*)$, and the posterior mean of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})$ can be estimated using the method in Section 2.1 by treating $\boldsymbol{\psi}$ as an additional block.

The BMI can also be used to assess the convergence of the Gibbs sampler, by computing and monitoring its stability for different iterations. Such an idea, combined with a different approach for computing the posterior density, appears in the Gibbs stopper proposed by Ritter and Tanner (1992). Raftery (1994) mentioned using the kernel estimate of the posterior density in connection with the BMI, but the resulting estimate can inherit the inaccuracy of the kernel method, especially in high dimensions. Finally, another identity similar to the BMI is available in the prediction context. Suppose that $y_f$ denotes an out-of-sample observation. Then the Bayesian prediction density, $f(y_f|\mathbf{y}) = \int f(y_f|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}$, can be expressed as

$$f(y_f|\mathbf{y}) = \frac{f(y_f|\mathbf{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y}, y_f)}$$

(see Besag 1989). This identity follows in a straightforward manner from the definition of the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}, y_f)$ and cross-multiplying. Besag (1989) alluded to a different proof.

## 3. NUMERICAL STANDARD ERROR

As mentioned in the preceding section, the proposed density estimation procedure is likely to produce an accurate estimate of $\pi(\boldsymbol{\theta}|\mathbf{y})$ at the point $\boldsymbol{\theta}^*$. In fact, it is possible to calculate the accuracy achieved by a computation that uses the Gibbs output. This calculation yields the numerical standard error of the marginal density estimate (or, equivalently, that of the posterior density estimate). The numerical standard error gives the variation that can be expected in the estimate if the simulation were to be done afresh, but the point at which the ordinate is evaluated is kept fixed.

To concentrate on the main ideas, consider the case in Section 2.1.2 and define the vector stochastic process

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

where in the first component the latent vector $(\boldsymbol{\theta}_2, \mathbf{z}) \sim [\cdot|\mathbf{y}]$ while in the second component the latent vector $\mathbf{z}$ follows the distribution $[\cdot|\mathbf{y}, \boldsymbol{\theta}_1^*]$. In general, $\mathbf{h}$ is a $B \times 1$ vector with the $r$th component given by $\pi(\boldsymbol{\theta}_r^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{r-1}^*, \boldsymbol{\theta}_l (l > r), \mathbf{z})$, the integrand of (10).

It should be noted that due to the procedure used to estimate the reduced conditional ordinate, the second component of $\mathbf{h}$ is approximately independent of the first. But for expositional simplifications, it is worthwhile to proceed with the vector formulation. Then in this notation,

$$\hat{\mathbf{h}} \equiv G^{-1}\sum_{g=1}^{G} \mathbf{h}^{(g)} = \begin{pmatrix} \hat{\pi}(\boldsymbol{\theta}_1^*|\mathbf{y}) \\ \hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) \end{pmatrix}, \tag{13}$$

and our objective is to find the variance of two functions of $\hat{\mathbf{h}}$, namely $\psi_1 = \hat{h}_1 \times \hat{h}_2$ and $\psi_2 = \ln(\hat{h}_1) + \ln(\hat{h}_2) \equiv \ln \hat{\pi}(\boldsymbol{\theta}_1^*|\mathbf{y}) + \ln \hat{\pi}(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*)$. The variance of these two functions is found by the delta method as soon as the variance of $\hat{\mathbf{h}}$ is determined. Because $\mathbf{h}$ inherits the ergodicity of the Gibbs output, it follows by the ergodic theorem (Tierney 1994) that

$$\hat{\mathbf{h}} \to \boldsymbol{\mu}, \quad \text{as} \quad G \to \infty,$$

almost surely, where $\boldsymbol{\mu} = (\pi(\boldsymbol{\theta}_1^*|\mathbf{y}), \pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*))'$,

$$\lim_{G\to\infty} G\{E(\hat{\mathbf{h}} - \boldsymbol{\mu})(\hat{\mathbf{h}} - \boldsymbol{\mu})'\} = 2\pi\mathbf{S}(0),$$

and $\mathbf{S}(0)$ is the spectral density matrix at frequency zero. An estimate of $\boldsymbol{\Omega} \equiv 2\pi\mathbf{S}(0)$ can be obtained by the approach of Newey and West (1987) or Geweke (1992). If

$$\boldsymbol{\Omega}_s = G^{-1}\sum_{g=s+1}^{G} (\mathbf{h}^{(g)} - \hat{\mathbf{h}})(\mathbf{h}^{(g)} - \hat{\mathbf{h}})',$$

then

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

Table 1. Nodal Involvement Data

| Case | y | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | Case | y | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 66 | .48 | 0 | 0 | 0 | 2 | 0 | 68 | .56 | 0 | 0 | 0 |
| 3 | 0 | 66 | .50 | 0 | 0 | 0 | 4 | 0 | 56 | .52 | 0 | 0 | 0 |
| 5 | 0 | 58 | .50 | 0 | 0 | 0 | 6 | 0 | 60 | .49 | 0 | 0 | 0 |
| 7 | 0 | 65 | .46 | 1 | 0 | 0 | 8 | 0 | 60 | .62 | 1 | 0 | 0 |
| 9 | 1 | 50 | .56 | 0 | 0 | 1 | 10 | 0 | 49 | .55 | 1 | 0 | 0 |
| 11 | 0 | 61 | .62 | 0 | 0 | 0 | 12 | 0 | 58 | .71 | 0 | 0 | 0 |
| 13 | 0 | 51 | .65 | 0 | 0 | 0 | 14 | 1 | 67 | .67 | 1 | 0 | 1 |
| 15 | 0 | 67 | .47 | 0 | 0 | 1 | 16 | 0 | 51 | .49 | 0 | 0 | 0 |
| 17 | 0 | 56 | .50 | 0 | 0 | 1 | 18 | 0 | 60 | .78 | 0 | 0 | 0 |
| 19 | 0 | 52 | .83 | 0 | 0 | 0 | 20 | 0 | 56 | .98 | 0 | 0 | 0 |
| 21 | 0 | 67 | .52 | 0 | 0 | 0 | 22 | 0 | 63 | .75 | 0 | 0 | 0 |
| 23 | 1 | 59 | .99 | 0 | 0 | 1 | 24 | 0 | 64 | 1.87 | 0 | 0 | 0 |
| 25 | 1 | 61 | 1.36 | 1 | 0 | 0 | 26 | 1 | 56 | .82 | 0 | 0 | 0 |
| 27 | 0 | 64 | .40 | 0 | 1 | 1 | 28 | 0 | 61 | .50 | 0 | 1 | 0 |
| 29 | 0 | 64 | .50 | 0 | 1 | 1 | 30 | 0 | 63 | .40 | 0 | 1 | 0 |
| 31 | 0 | 52 | .55 | 0 | 1 | 1 | 32 | 0 | 66 | .59 | 0 | 1 | 1 |
| 33 | 1 | 58 | .48 | 1 | 1 | 0 | 34 | 1 | 57 | .51 | 1 | 1 | 1 |
| 35 | 1 | 65 | .49 | 0 | 1 | 0 | 36 | 0 | 65 | .48 | 0 | 1 | 1 |
| 37 | 0 | 59 | .63 | 1 | 1 | 1 | 38 | 0 | 61 | 1.02 | 0 | 1 | 0 |
| 39 | 0 | 53 | .76 | 0 | 1 | 0 | 40 | 0 | 67 | .95 | 0 | 1 | 0 |
| 41 | 0 | 53 | .66 | 0 | 1 | 1 | 42 | 1 | 65 | .84 | 1 | 1 | 1 |
| 43 | 1 | 50 | .81 | 1 | 1 | 1 | 44 | 1 | 60 | .76 | 1 | 1 | 1 |
| 45 | 1 | 45 | .70 | 0 | 1 | 1 | 46 | 1 | 56 | .78 | 1 | 1 | 1 |
| 47 | 1 | 46 | .70 | 0 | 1 | 0 | 48 | 1 | 67 | .67 | 0 | 1 | 0 |
| 49 | 1 | 63 | .82 | 0 | 1 | 0 | 50 | 1 | 57 | .67 | 0 | 1 | 1 |
| 51 | 1 | 51 | .72 | 1 | 1 | 0 | 52 | 1 | 64 | .89 | 1 | 1 | 0 |
| 53 | 1 | 68 | 1.26 | 1 | 1 | 1 | | | | | | | |

where $q$ is some constant, essentially the value at which the autocorrelation function tapers off. In the applications to follow $q$ is conservatively set equal to 10, although there was negligible to vanishing serial correlation in the $\hat{\mathbf{h}}^{(g)}$ process. The variance of $\psi_2$, for example, is found by the delta method to be

$$\left(\frac{\partial \psi_2}{\partial \hat{\mathbf{h}}}\right)' \text{var}(\hat{\mathbf{h}}) \left(\frac{\partial \psi_2}{\partial \hat{\mathbf{h}}}\right), \tag{14}$$

where the derivative vector consists of elements $\hat{h}_1^{-1}$ and $\hat{h}_2^{-1}$. The square root of this variance is the numerical standard error of the marginal likelihood in the log scale.

## 4. EXAMPLES

In this section the approach developed earlier is applied to two important classes of models. In particular, the methods are discussed in the context of variable selection in binary probit regression models and in the context of two broad classes of finite mixture models, the iid mixture model and the Markov mixture model.

By way of notation, for a $d$-dimensional normal random vector with mean $\mu$ and covariance matrix $\mathbf{\Sigma}$, the density at the point $\mathbf{t}$ is denoted by $\phi(\mathbf{t}|\boldsymbol{\mu}, \mathbf{\Sigma}) \equiv (2\pi)^{d/2}|\mathbf{\Sigma}|^{-1/2}\exp(-(\mathbf{t}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})/2)$ and the inverse gamma density at the point $s$ is denoted by $p_{IG}(s|a,b) \equiv (b^a/\Gamma(a))(1/s)^{(a+1)}\exp(-b/s)$. Finally, for a $m$ vector $\mathbf{q}$ on the unit simplex, the Dirichlet $D(a_1, \alpha_2, \ldots, \alpha_m)$ density is denoted by $p_D(\mathbf{q}|\alpha_1, \ldots, \alpha_m) \equiv \Gamma(\sum_j \alpha_j) q_1^{\alpha_1 - 1} \cdots q_m^{\alpha_k - 1} / \prod_j \Gamma(\alpha_j)$.

### 4.1 Binary Probit Regression

Consider the data in Table 1 on the presence of prostatic nodal involvement collected on 53 patients with cancer of the prostate. The data (reported in the study by Brown (1980); see also Collett 1991) include a binary response variable $y$ that takes the value 1 if cancer had spread to the surrounding lymph nodes and value zero otherwise. The objective is to explain the binary response with five variables: age of the patient in years at diagnosis ($x_1$); level of serum acid phosphate ($x_2$); the result of an X-ray examination, coded 0 if negative and 1 if positive ($x_3$); the size of the tumor, coded 0 if small and 1 if large ($x_4$); and the pathological grade of the tumor, coded 0 if less serious and 1 if more serious ($x_5$).

The probability of positive response can be explained through a probit link function or, as by Collett (1991), by a logit link. If interactions and powers of explanatory variables are excluded, then there are 32 possible models that can be fit. Collett's finding from the classical deviance statistic ($-2$ times the maximized log-likelihood) is that the logistic model containing $\log(x_2), x_3$, and $x_4$ provides a suitable fit for the data among these 32 models. These data are reanalyzed to demonstrate the computation of the marginal likelihood using nine of these models (defined later and selected entirely for illustrative purposes).

Under model $k$, suppose that

$$\Pr(y_i = 1|M_k) = \Phi(\mathbf{x}_{ik}'\boldsymbol{\beta}_k), \qquad i \leq 53,$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal density, $x_{ik}$ are the covariates included in model $k$, and $\boldsymbol{\beta}_k$ is the corresponding regression parameter vector. The likelihood function under $M_k$, assuming a

Table 2. Summary of Results for Nodal Involvement Data

| Terms fitted in model | log (maximized lik) | d.f. | log (marginal) | Num SE |
|---|---|---|---|---|
| $C$ | −35.126 | 52 | −38.503 | .005 |
| $C + x_1$ | −34.587 | 51 | −43.175 | .007 |
| $C + \log(x_2)$ | −32.431 | 51 | −37.916 | .007 |
| $C + x_3$ | −29.500 | 51 | −35.323 | .009 |
| $C + x_4$ | −31.276 | 51 | −37.234 | .009 |
| $C + x_5$ | −33.099 | 51 | −39.075 | .007 |
| $C + \log(x_2) + x_4$ | −28.187 | 50 | −36.140 | .013 |
| $C + \log(x_2) + x_3 + x_4$ | −24.427 | 49 | −34.553 | .020 |
| $C + \log(x_2) + x_3 + x_4 + x_5$ | −23.769 | 48 | −36.233 | .024 |

random sample, is then

$$f(\mathbf{y}|M_k, \boldsymbol{\beta}_k) = \prod_{i=1}^{53} [\Phi(\mathbf{x}_{ik}'\boldsymbol{\beta}_k)]^{y_i}[1 - \Phi(\mathbf{x}_{ik}'\boldsymbol{\beta}_k)]^{1-y_i}.$$

For this situation, the marginal likelihood can be computed rather simply by the Laplace method (see Kass and Raftery 1994), but given the small sample size, it is difficult to know the accuracy of the Laplace approximation. Harmonic mean type estimators, on the other hand, are rather more difficult to obtain with this likelihood, because its tails generally decline quite sharply.

A procedure that works extremely well in conjunction with the technique developed above is the data augmentation–Gibbs sampling method of Albert and Chib (1993a). Suppose that the prior information about $\boldsymbol{\beta}_k$ is weak, but not improper, and is represented by a multivariate normal prior with the mean of each parameter equal to .75 (because each covariate is expected to have a positive impact on the probability of response), and a standard deviation of 5. Under the assumption that the parameters are independent, the prior of $\boldsymbol{\beta}_k$ takes the form

$$\boldsymbol{\beta}_k \sim N(\mathbf{a}_k, \mathbf{A}_k^{-1}).$$

Suppressing the model index $k$, the Gibbs draws for each model are obtained as follows. Define a normally distributed latent variable, $z_i$, such that

$$z_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, 1); \qquad y_i = I(z_i > 0),$$

where $I(A)$ is an indicator function of the event $A$. This in fact is equivalent to the probit model, because $\Pr(z_i > 0) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$. Then, following Albert and Chib (1993a), the Gibbs sampler is defined through the complete conditional densities

$$\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{z}) = \phi(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_\mathbf{z}, \mathbf{B})$$

and

$$\begin{aligned} p(z_i|\mathbf{y}, \boldsymbol{\beta}) &\propto \phi(z_i|\mathbf{x}_i'\boldsymbol{\beta}, 1)I[0, \infty] \quad \text{if } y_i = 1, \\ &\propto \phi(z_i|\mathbf{x}_i'\boldsymbol{\beta}, 1)I[-\infty, 0] \quad \text{if } y_i = 0, \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_\mathbf{z} = (\mathbf{A} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{A}\mathbf{a} + \mathbf{X}'\mathbf{z}), \mathbf{B} = (\mathbf{A} + \mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$, $\mathbf{X}$ is the matrix of all the covariates, and $\phi(\cdot|\mu, 1)I[a, b]$ is the normal density truncated to the interval $[a, b]$. The output $\{\boldsymbol{\beta}^{(g)}, \hat{\boldsymbol{\beta}}_\mathbf{z}^{(g)}\}$ is obtained by running the sampler for $G = 5{,}000$ cycles after deleting the first 500, and the estimate $\boldsymbol{\beta}^* = \sum \boldsymbol{\beta}^{(g)}/5{,}000$ is obtained.

Then the logarithm of the marginal likelihood of model $M_k$ is

$$\begin{aligned} \ln f(\mathbf{y}|M_k, \boldsymbol{\beta}_k^*) \quad &+ \quad \ln \phi(\boldsymbol{\beta}_k^*|\mathbf{a}_k, \mathbf{A}_k^{-1}) \\ &- \quad \ln \left\{ 5{,}000^{-1} \sum_{g=1}^{5{,}000} \phi(\boldsymbol{\beta}_k^*|\hat{\boldsymbol{\beta}}_\mathbf{z}^{(g)}, \mathbf{B}_k) \right\}, \quad (15) \end{aligned}$$

where, it should be noted, the mean vector of the third density (i.e., $\hat{\boldsymbol{\beta}}_\mathbf{z}^{(g)}$) is produced as a by-product of the sampling algorithm.

The results are summarized in Table 2, where for each of nine models, the maximized likelihood is reported along with the degrees of freedom, the log of the marginal likelihood, and its numerical standard error. From this table it can be seen that the marginal likelihood is very precisely estimated in all the fitted models. Of course, these results are obtained with $G = 5{,}000$ draws, and further improvements in accuracy can be achieved by increasing $G$. For comparison, the BMI expression was evaluated at a point that was one posterior standard deviation from $\boldsymbol{\beta}^*$. As expected, this led to an increase in the numerical standard error of the estimate and, for example, was .26 in $M_9$, with $G = 5{,}000$. The Laplace method was also used to determine the marginal likelihood, and the results were in agreement up to the second decimal place. We also examined if a multivariate kernel estimate of the posterior ordinate (with a Gaussian product kernel) could be used in the BMI expression. This procedure did not produce equally accurate results. Also note that $x_1$ (the age variable) does not improve on the model with just a constant (the Bayes factor for the second model vs. the first is .009), whereas the model with the variable $x_3$ (X-ray) has a Bayes factor of approximately 25 versus the model with just a constant. The Bayes factor for $M_8$ versus $M_9$ is 5.33, supporting the conclusion of Collett (1991), who argued that $M_8$ is the best model, and also demonstrating the value of the marginal likelihood in providing information about the comparative value of a fitted model.

### 4.2 Marginal Likelihood in Mixture Models

To further illustrate the usefulness of our approach, con-

Table 3. Velocity (km/second) for Galaxies in the Corona Borealis Region

| | | | | | |
|---|---|---|---|---|---|
| 9,172 | 9,350 | 9,483 | 9,558 | 9,775 | 10,227 |
| 10,406 | 16,084 | 16,170 | 18,419 | 18,552 | 18,600 |
| 18,927 | 19,052 | 19,070 | 19,330 | 19,343 | 19,349 |
| 19,440 | 19,473 | 19,529 | 19,541 | 19,547 | 19,663 |
| 19,846 | 19,856 | 19,863 | 19,914 | 19,918 | 19,973 |
| 19,989 | 20,166 | 20,175 | 20,179 | 20,196 | 20,215 |
| 20,221 | 20,415 | 20,629 | 20,795 | 20,821 | 20,846 |
| 20,875 | 20,986 | 21,137 | 21,492 | 21,701 | 21,814 |
| 21,921 | 21,960 | 22,185 | 22,209 | 22,242 | 22,249 |
| 22,314 | 22,374 | 22,495 | 22,746 | 22,747 | 22,888 |
| 22,914 | 23,206 | 23,241 | 23,263 | 23,484 | 23,538 |
| 23,542 | 23,666 | 23,706 | 23,711 | 24,129 | 24,285 |
| 24,289 | 24,366 | 24,717 | 24,990 | 25,633 | 26,960 |
| 26,995 | 32,065 | 32,789 | 34,279 | | |

*Table 4. Summary of Results for Galaxy Data*

| Model fitted | log(marginal) | Num SE |
|---|---|---|
| Two components: $\sigma_j^2 = \sigma^2, \forall\, j$ | −240.464 | .006 |
| Three components: $\sigma_j^2 = \sigma^2, \forall\, j$ | −228.620 | .008 |
| Three components: $\sigma_j^2$ unrestricted | −224.138 | .086 |

sider the calculation of the marginal likelihood in two broad applications that involve mixture models. The first is concerned with determining the number of components in a Gaussian finite mixture model applied to astronomical data on the velocity of galaxies. The second is concerned with a mixture model that applies to time series data. This model, which is also referred to as a Markov switching model or a hidden Markov model, is illustrated with data on the growth rates of U.S. gross national product for the postwar period.

*4.2.1 Determining the Number of Components in a Mixture.* Consider the data set in Table 3 on velocities of 82 galaxies from 6 well-separated conic sections of the Corona Borealis region, originally presented by Postman, Huchra, and Geller (1986). The objective is to find the best-fitting Gaussian finite mixture model. This data set has been analyzed by Roeder (1990) who developed a nonparametric density approach to determine the number of modes. Subsequently, Carlin and Chib (1993) reanalyzed the data by parametric Bayesian methods and estimated Gaussian mixture models with two to five components using the Gibbs sampler. Their results indicate symptoms of overfitting when models with four or five components are estimated. The Gibbs output from these models displays nonvanishing serial correlation for extremely high lags, indicating difficulties with convergence and nonidentifiability of parameters. (See Crawford 1994 for a discussion of identification issues in mixture models.) For this reason, models with two and three components are fit.

For the model with $d$ components, suppose that the $j$th component is given by $\phi(y_i|\mu_j, \sigma_j^2)$, where $y_i$ is $i$th data value (velocity/1,000) and $(\mu_j, \sigma_j^2)$ is the component-specific mean and variance. If each component is sampled with probability $q_j (\sum q_j = 1)$, then the density function of the data $\mathbf{y} = (y_1, \ldots, y_{82})$ given the parameters $\boldsymbol{\theta}$ is

$$f(\mathbf{y}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{d} q_j \phi(y_i|\mu_j, \sigma_j^2), \tag{16}$$

where $\boldsymbol{\theta} = (\mathbf{q}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ with $\mathbf{q} = (q_1, q_2, \ldots, q_d)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_d^2)$. It is useful to refer to this model as the "iid mixture model" because, as is well known, by introducing iid latent variables $z_i \in \{1, 2, \ldots, d\}$ such that

$$\Pr(z_i = j) = q_j \tag{17}$$

and defining $f(y_i|z_i = j, \boldsymbol{\theta}) = \phi(y_i|\mu_j, \sigma_j^2)$ leads to the mixture model in (16).

Assume that all components of $\boldsymbol{\theta}$ are mutually independent, and define the prior information through the distributions

$$\mu_j \sim \mathcal{N}(\mu_0, A^{-1}),$$

$$\sigma_j^2 \sim \mathcal{IG}\left(\frac{v_0}{2}, \frac{\delta_0}{2}\right),$$

and

$$\mathbf{q} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d),$$

where $\mu_0 = 20, A^{-1} = 100, v_0 = 6, \delta_0 = 40$, and $\alpha_j = 1$. As can be observed, these priors reflect weak prior information about the parameters. Under these prior distributions, the objective is to compute the marginal likelihood for models with two and three components. In addition, models obtained by restricting the variance $\sigma_j^2$ to be constant across components are also of interest.

The Gibbs implementation for this model is straightforward (see Diebolt and Robert 1994 and West 1992). Let $\mathbf{z} = (z_1, \ldots, z_n)$, then Gibbs sampling is defined through the conditional densities of $\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{q}$, and $\mathbf{z}$. Let $T_j = \{i : z_i = j\}$ be the set of observation indices for the observations classified into the $j$th population and let $n_j$ represent the number of observations so assigned. Now pick out the observations that correspond to the $j$th population and place them in the vector $\mathbf{y}_j$ and define an $n_j$ vector $\mathbf{i}_j$ comprising of units. Then

$$\pi(\boldsymbol{\mu}|y, \mathbf{z}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{d} \phi(\mu_j|\hat{\mu}_j, B_j),$$

$$\pi(\boldsymbol{\sigma}^2|\mathbf{y}, \mathbf{z}, \boldsymbol{\mu}) = \prod_{j=1}^{d} p_{IG}(\sigma_j^2|\{v_0 + n_j\}/2, \{\delta_0 + \delta_j\}/2),$$

and

$$\pi(\mathbf{q}|\mathbf{y}, \mathbf{z}) = p_D(\mathbf{q}|\alpha_1 + n_1, \ldots, \alpha_d + n_d),$$

and $\Pr(z_i = j|\mathbf{y}, \theta) \propto q_j \times \phi(y_i|\mu_j, \sigma^2), i \leq n$, where $\hat{\mu}_j = (A + \sigma_j^{-2} n_j)^{-1}(A\mu_0 + \sigma_j^{-2}\mathbf{i}_j'\mathbf{y}_j), B_j = (A + \sigma_j^{-2} n_j)^{-1}$, and $\delta_j = (\mathbf{y}_j - \mathbf{i}_j\mu_j)'(\mathbf{y}_j - \mathbf{i}_j\mu_j)$.

The posterior density ordinate can be computed from the decomposition

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \pi(\boldsymbol{\mu}^*|\mathbf{y}) \times \pi(\boldsymbol{\sigma}^{2*}|\boldsymbol{\mu}^*, \mathbf{y}) \times \pi(\mathbf{q}^*|\mathbf{y}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}), \tag{18}$$

where $\boldsymbol{\theta}^*$ is taken to be the (approximate) maximum likelihood estimate computed by evaluating (16) for each simulated draw. Now apply (11) as follows:

- The draws from the full Gibbs run are used to estimate $\pi(\boldsymbol{\mu}^*|\mathbf{y}) = \int \prod_{j=1}^{d} \phi(\mu_j|\hat{\mu}_j, B_j)\pi(\mathbf{z}, \boldsymbol{\sigma}^2|\mathbf{y})\, d\mathbf{z}\, d\boldsymbol{\sigma}^2$.
- Next, the draws from the reduced Gibbs run with the densities $\pi(\sigma_j^2|\mathbf{y}, \mathbf{z}, \boldsymbol{\mu}^*), \pi(\mathbf{q}|\mathbf{y}, \mathbf{z})$ and $\{\Pr(z_i|\mathbf{y}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^2, \mathbf{q})\}$ are used to estimate $\pi(\boldsymbol{\sigma}^{2*}|\boldsymbol{\mu}^*, \mathbf{y}) = \int \prod_{j=1}^{d} p_{IG}(\sigma_j^2|\frac{1}{2}\{v_0 + n_j\}, \frac{1}{2}\{\delta_0 + \delta_j\})\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\mu}^*)\, d\mathbf{z}$.
- Finally, the draws from the subsequent reduced Gibbs run with the densities $\pi(\mathbf{q}|\mathbf{y}, \mathbf{z})$ and $\{\Pr(z_i|\mathbf{y}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}, \mathbf{q})\}$ are used to estimate $\pi(\mathbf{q}^*|\mathbf{y}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}) = \int p_D(\mathbf{q}|\alpha_1 + n_1, \ldots, \alpha_d + n_d)p(\mathbf{z}|\mathbf{y}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*})\, d\mathbf{z}$.

Table 5. Summary of Results for U.S. GNP Growth Rates Data

| | Prior | | Posterior | |
|---|---|---|---|---|
| Parameter | Mean | Std dev | Mean | Std dev |
| $\mu_1$ | 0 | 1.414 | − .313 | .314 |
| $\mu_2$ | .75 | 1.414 | 1.038 | .111 |
| $\sigma_2$ | 1.33 | .943 | .672 | .089 |
| $p_{11}$ | .8 | .163 | .743 | .098 |
| $p_{22}$ | .8 | .163 | .911 | .042 |
| | Log marginal likelihood | | −229.496 (.028) | |

An estimate of the marginal likelihood is given by substituting these quantities into (12).

Our results, which are based on $G = 5{,}000$ draws, are summarized in Table 4. (Almost identical results were obtained when the BMI expression was evaluated at the posterior mean instead of the approximate maximum likelihood value.) First, the two-component model is clearly dominated by both three-component models. Second, the three-component model with $\sigma^2$ unrestricted appears to be better than the three-component model with $\sigma^2$ restricted to be the same across components. This result would not be obvious from just looking at posterior distributions of the fitted models, because all the parameters in both three-component models are tightly estimated. Third, all the numerical standard errors are small, indicating that the marginal likelihood has been accurately estimated.

*4.2.2 Markov Mixture Model.* As a final illustration of the value of our approach, consider data on the quarterly growth rates of U.S. gross national product (GNP) for the postwar period 1951.2 to 1992.4. Many different time series models have been fit to this data, and our objective is to demonstrate how the marginal likelihood can be calculated in one particular case, of substantial practical importance, for which this calculation has hitherto not been attempted.

The model of interest is the Markov mixture model, also sometimes referred to as the Markov switching model (Goldfeld and Quandt 1973; Hamilton 1989). Let $y_t$ denote the growth rate of GNP (multiplied by 100), and suppose that

$$y_t|\boldsymbol{\mu}, z_t = j, \sigma^2 \sim \mathcal{N}(\mu_j, \sigma^2), \qquad j = 1, 2,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $z_t$ is an unobserved state variable that follows a two-state Markov chain,

$$z_t|z_{t-1}, \mathbf{P} \sim \text{Markov}(\mathbf{P}, \boldsymbol{\pi}_1), \tag{19}$$

where $\mathbf{P} = \{p_{ij}\}$ is the one-step transition probability matrix of the chain (i.e., $p_{ij} = \Pr(z_t = j|z_{t-1} = i)$, and $\boldsymbol{\pi}_1$ is the probability distribution at $t = 1$. This model is a generalization of the iid mixture model of the last subsection. Furthermore, it is a model that is particularly appropriate for modeling correlation in growth rates that are observed in practice.

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \mathbf{q}_1, \mathbf{q}_2)$, where $\mathbf{q}_i$ is the $i$th row of $\mathbf{P}$; then the likelihood function for the Markov mixture model is given in terms of the one-step ahead prediction densities,

$$f(y_t|Y_{t-1}, \boldsymbol{\theta}) = p(z_t = 1|Y_{t-1}, \boldsymbol{\theta})\phi(y_t|\mu_1, \sigma^2) + (1 - p(z_t = 1|Y_{t-1}, \boldsymbol{\theta}))\phi(y_t|\mu_2, \sigma^2),$$

where $Y_{t-1}$ is the observed data up to time $t - 1$ and $p(z_t = 1|Y_{t-1}, \boldsymbol{\theta})$ is a time-varying conditional probability. The joint density of all the data is then

$$f(\mathbf{y}|\theta) = \prod_{t=1}^{n} f(y_t|Y_{t-1}, \boldsymbol{\theta}). \tag{20}$$

A little reflection shows that, given $\mathbf{z}$, this model has the same structure as the iid mixture model, and thus the marginal likelihood calculation proceeds in virtually the same way. The complete conditional densities of $(\boldsymbol{\mu}, \sigma^2)$ are identical to those in the iid mixture model, and, if one assumes that the prior density on $\mathbf{q}_i$ is Dirichlet$(\alpha_{i1}, \alpha_{i2})$, then

$$\pi(\mathbf{q}_1, \mathbf{q}_2|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^{2} p_D(\mathbf{q}_i|\alpha_{i1} + n_{i1}, \alpha_{12} + n_{i2}),$$

where $n_{ij}$ denotes the number of one-step transitions from $i$ to $j$ in the sequence $\mathbf{z}$ (see Albert and Chib 1993b). A decomposition similar to (18) is again available while each of the ordinates can be estimated by the reduced conditional Gibbs sampling procedure described earlier.

The Gibbs implementation of this model, and the calculation of the marginal likelihood, require the simulation of the latent variables $\mathbf{z}$ from $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$. As described by Chib (1993), the latent variables are simulated through the following recursive steps, which are initiated with $p(z_0 = i|Y_0, \boldsymbol{\theta})$. These recursions require one pass from $t = 1$ to $n$ and then a second pass from $t = n$ to $t = 1$.

Step 1: Repeat for $t = 1, 2, \ldots, n$.

Prediction step: Calculate

$$p(z_t = j|Y_{t-1}, \boldsymbol{\theta}) = \sum_{i=1}^{2} p_{ij} \times p(z_{t-1} = i|Y_{t-1}, \boldsymbol{\theta}), \qquad (j = 1, 2).$$

Update step: Calculate

$$p(z_t = j|Y_t, \boldsymbol{\theta}) \propto p(z_t = j|Y_{t-1}, \boldsymbol{\theta}) \times \phi(y_t|\mu_j, \sigma^2).$$

Step 2: Simulate $z_n$ from $p(z_n = j|Y_n, \boldsymbol{\theta})$, the mass function produced by the last update step.

Step 3: Repeat for $t = n - 1, \ldots, 2, 1$.

Given the draw $z_{t+1} = l$, calculate

$$p(z_t = j|Y_n, z_{t+1} = l, \boldsymbol{\theta}) \propto p_{jl} \times p(z_t = j|Y_t, \boldsymbol{\theta}), \qquad (j = 1, 2).$$

Simulate $z_t$ from $p(z_t = j|Y_n, z_{t+1} = l, \boldsymbol{\theta})$.

Note that the prediction step gives the time-varying probability mass function required to calculate the likelihood function in (20) at the point $\boldsymbol{\theta}^*$.

Our results for this model and data are summarized in Table 5. These results are based on $G = 6{,}000$ draws and rely on the prior distributions $\mu_1 \sim \mathcal{N}(0, 2), \mu_2$

$\sim \mathcal{N}(.75, 2), \sigma^2 \sim \mathcal{IG}(4,4), \mathbf{q}_1 \sim$ Dirichlet$(4,1)$, and $\mathbf{q}_2 \sim$ Dirichlet$(1,4)$. These priors are relatively vague and are designed to model the potential persistence in low and high growth rates. Thus the marginal likelihood is seen to be equal to $-229.496$ on the log scale and is accurately estimated with a numerical standard error of .028. In comparison, the marginal likelihood is also calculated for a first-order autoregressive model $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t, \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, by treating this as a linear regression model after conditioning on the first observation. Under the prior $(\beta_1, \beta_2)' \sim \mathcal{N}_2(0, \text{diag}(10,10))$ and $\sigma^2 \sim \mathcal{IG}(3,3)$, the log marginal likelihood is estimated to be $-231.94$. Thus the data support the Markov mixture model to the first-order autoregressive model.

## 5. CONCLUDING REMARKS

In summary, this article has developed and illustrated a new approach to calculating the marginal likelihood that relies on the output of the Gibbs sampling algorithm. The approach is fully automatic and stable, requiring no inputs beyond the draws from the simulation. Thus draws from the prior, or additional maximizations, or importance sampling functions, or any other tuning function, are not required. It was shown that the numerical standard error of the estimate can be derived from the posterior sample and the calculations are exhibited in problems dealing with probit regression and finite-mixture models. In all the examples, the marginal likelihood is estimated easily and very accurately. As a result, this approach should encourage the routine calculation of Bayes factors in models estimated by the Gibbs sampler.

*[Received May 1994. Revised February 1995.]*

## REFERENCES


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

——— (1993b), "Bayes Inference Via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *Journal of Business & Economic Statistics,* 11, 1–15.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis,* New York: Springer-Verlag.

Besag, J. (1989), "A Candidates Formula: A Curious Result in Bayesian Prediction," *Biometrika,* 76, 183.

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: John Wiley.

Carlin, B., and Chib, S. (1993), "Bayesian Model Choice Via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society,* Ser. B, 57, 473–484.

Carlin, B., and Polson, N. (1991), "Inference for Nonconjugate Bayesian Models Using Gibbs Sampling," *Canadian Journal of Statistics,* 19, 399–405.

Chib, S. (1992), "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics,* 51, 79–99.

——— (1993), "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," submitted to *Journal of Econometrics.*

Collett, D. (1991), *Modelling Binary Data,* London: Chapman and Hall.

Crawford, S. L. (1994), "An Application of the Laplace Method to Finite Mixture Distributions," *Journal of the American Statistical Association,* 89, 259–267.

Diebolt, J., and Robert, C. P. (1993), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society,* Ser. B, 56, 363–375.

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

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.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 169–193.

Goldfeld, S. M., and Quandt, R. E. (1973), "A Markov Model for Switching Regressions," *Journal of Econometrics,* 1, 3–16.

Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica,* 57, 357–384.

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

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

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

O'Hagan, A. (1994), *Bayesian Inference* (Kendall's Advanced Theory of Statistics, Vol. 2B), London: Edward Arnold.

Postman, M., Huchra, J. P., and Geller, M. J. (1986), "Probes of Large-Scale Structures in the Corona Borealis Region," *The Astronomical Journal,* 92, 1238–1247.

Raftery, A. E. (1994), "Hypothesis Testing and Model Selection Via Posterior Simulation," unpublished manuscript, University of Washington, Dept. of Statistics.

Roeder, K. (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in Galaxies," *Journal of the American Statistical Association,* 85, 617–624.

Scott, D. W. (1992), *Multivariate Density Estimation,* New York: John Wiley.

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. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

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

West, M. (1992), "Modelling With Mixtures" (with discussion), in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 503–524.

Zellner, A., and Min, C. (1995), "Gibbs Sampler Convergence Criteria (GSC$^2$)," *Journal of the American Statistical Association,* 90, 921–927.