Statistica Neerlandica (2005) Vol. 59, nr. 1, pp. 30–44

# Accept–reject Metropolis–Hastings sampling and marginal likelihood estimation

Siddhartha Chib*

*John M. Olin School of Business, Washington University,*
*Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130*

Ivan Jeliazkov†

*Department of Economics, University of California, Irvine,*
*3151 Social Science Plaza, Irvine, CA 92697-5100*

We describe a method for estimating the marginal likelihood, based on Chib (1995) and Chib and Jeliazkov (2001), when simulation from the posterior distribution of the model parameters is by the accept–reject Metropolis–Hastings (ARMH) algorithm. The method is developed for one-block and multiple-block ARMH algorithms and does not require the (typically) unknown normalizing constant of the proposal density. The problem of calculating the numerical standard error of the estimates is also considered and a procedure based on batch means is developed. Two examples, dealing with a multinomial logit model and a Gaussian regression model with non-conjugate priors, are provided to illustrate the efficiency and applicability of the method.

*Key Words and Phrases:* Model comparison, Bayes factor, Gaussian regression, lognormal density, log-*t* density, Markov chain Monte Carlo, logit model.

## 1 Introduction

In this article we describe a method for estimating the marginal likelihood of a model, for the purpose of comparing models via Bayes factors, from the building blocks of the accept–reject Metropolis–Hastings (ARMH) algorithm (Tierney, 1994; Chib and Greenberg, 1995). The method is based on the framework of Chib (1995), which has been widely used to estimate the marginal likelihood of Bayesian models from the output of Markov chain Monte Carlo (MCMC) simulations. Chib and Jeliazkov (2001) present a useful version of this approach for the case where some of the full-conditional densities are non-standard and sampling requires the use of the Metropolis–Hastings (M–H) algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995).

*chib@wustl.edu
†ivan@uci.edu

When MCMC simulation is implemented by the ARMH algorithm, one interesting challenge is that the normalizing constant of the M–H proposal density is typically unknown, as it depends on the target density of interest. A second difficulty arises in determining the variability of the marginal likelihood estimate, which utilizes draws from both the accept–reject (A–R) and the M–H part of the ARMH algorithm. While the A–R and M–H sequences are mutually dependent by construction, the dependence is complicated since A–R draws can be rejected and M–H draws can be repeated. Moreover, the two sequences are of unequal lengths – if one of the simulation sizes is fixed, the other is randomly determined. Here we show how these obstacles can be overcome to produce estimates that are efficient and economical in terms of programming, additional tuning effort, and computational intensity.

The proposed method joins a substantial literature concerned with the estimation of marginal likelihoods and Bayes factors (e.g. Newton and Raftery, 1994; Gelfand and Dey, 1994; Carlin and Chib, 1995; Chib, 1995; Green, 1995; Verdinelli and Wasserman, 1995; Meng and Wong, 1996; Chen and Shao, 1997; DiCiccio *et al.*, 1997; Chib and Jeliazkov, 2001; Basu and Chib, 2003). Han and Carlin (2001) offer a recent comparative review of some of these methods, in which they consider features such as computational simplicity, efficiency, and the additional overhead due to tuning and convergence concerns. In line with the procedure developed in Chib (1995), the approach proposed here reduces the implementation costs by estimating the marginal likelihood from the components of the sampling algorithm without requiring additional inputs (e.g. auxiliary densities or asymptotic approximations). Thus, once the coding of the simulation algorithm is completed, estimation of the marginal likelihood is conceptually straightforward.

The proposed techniques are illustrated in two examples involving logistic and Gaussian regression models. The first example considers data on commuters' work trips from Small (1982), while the second deals with data on women's wages from Mroz (1987). The examples provide practical evidence on the performance of the estimation and model choice methods under different ARMH designs.

The rest of the paper is organized as follows. Section 2 outlines the marginal likelihood estimation framework of Chib (1995) and Section 3 presents the ARMH algorithm. Section 4 contains our main results on the estimation of the marginal likelihood and its numerical standard error. We present two applications in Section 5, and concluding remarks in Section 6.

## 2 Preliminaries

The marginal likelihood, $m(y) \equiv \int f(y|\theta)\pi(\theta)\,\mathrm{d}\theta$, where $f(y|\theta)$ is the sampling density of the data $y$ and $\pi(\theta)$ is the prior density of the model parameters $\theta$, is of fundamental importance in Bayesian model comparison, because of its role in determining the posterior model probability. Specifically, the posterior odds of any two models are given by the prior odds of the models times their Bayes factor, defined as the ratio of their marginal likelihoods (Jeffreys, 1961). Chib (1995) provides a method for estimating the marginal likelihood that amounts to finding an estimate of the posterior density $\pi(\theta|y)$ at a single point in its support $\Theta$, by using the fact that $m(y)$ is the normalizing constant of the posterior density and hence can be found through the expression

$$m(y) = \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)}, \tag{1}$$

which follows from Bayes' formula, and is referred to as the basic marginal likelihood identity. Evaluating this expression on the log scale at some specific point $\theta^*$, one obtains $\log m(y) = \log f(y|\theta^*) + \log \pi(\theta^*) - \log \pi(\theta^*|y)$, where the first two terms on the right hand side are usually available by direct calculation. An estimate of the marginal likelihood, therefore, requires simply an estimate of the posterior ordinate $\pi(\theta^*|y)$. For estimation efficiency, the point $\theta^*$ is chosen in a high density region of the support $\Theta$.

Suppose that the parameters in a general MCMC sampler are grouped into $B$ blocks $\theta = (\theta_1,\ldots,\theta_B)$, with $\theta_k \in \Theta_k \subseteq \Re^{d_k}$, $k = 1,\ldots,B$. To facilitate the notation, let $\psi_i = (\theta_1,\ldots,\theta_i)$ denote the blocks up to $i$ and $\psi^{i+1} = (\theta_{i+1},\ldots,\theta_B)$ denote the blocks beyond $i$, and write the posterior ordinate at $\theta^*$ as

$$\pi(\theta_1^*,\ldots,\theta_B^*|y) = \prod_{i=1}^{B} \pi(\theta_i^*|y,\theta_1^*,\ldots,\theta_{i-1}^*) = \prod_{i=1}^{B} \pi(\theta_i^*|y,\psi_{i-1}^*). \tag{2}$$

Consider the estimation of a typical reduced ordinate $\pi(\theta_i^*|y,\ \psi_{i-1}^*)$. In the context of Gibbs sampling when the full-conditional densities, including their normalizing constants, are fully known, Chib (1995) proposed finding the ordinate $\pi(\theta_i^*|y,\ \psi_{i-1}^*)$ by Rao–Blackwellization

$$\begin{aligned}\pi(\theta_i^*|y,\psi_{i-1}^*) &= \int \pi(\theta_i^*|y,\psi_{i-1}^*,\psi^{i+1})\pi(\psi^{i+1}|y,\psi_{i-1}^*)\mathrm{d}\psi^{i+1}\\ &\approx G^{-1}\sum_{g=1}^{G}\pi\left(\theta_i^*|y,\psi_{i-1}^*,\psi^{i+1,(g)}\right),\end{aligned}$$

where $\psi^{i+1,(g)} \sim \pi(\psi^{i+1}|y,\ \psi_{i-1}^*)$ come from a *reduced run*, where the blocks $\psi_{i-1}^*$ are held fixed and sampling is only over $\psi^i$ (so that $\psi^{i+1,(g)}$ results by leaving out $\theta_i^{(g)}$). The ordinate $\pi(\theta_1^*|y)$ for the first block $\theta_1$ is estimated with draws $\theta \sim \pi(\theta|y)$ from the main MCMC run, while the ordinate $\pi(\theta_B^*|y,\ \psi_{B-1}^*)$ is available directly.

When one or more of the full-conditional densities are not of standard form and have intractable normalizing constants, posterior sampling is usually conducted via the M–H algorithm. In this case, Chib and Jeliazkov (2001) use the local reversibility of the M–H Markov chain to show that

$$\pi(\theta_i^*|y,\psi_{i-1}^*) = \frac{E_1\{\alpha_{MH}(\theta_i,\theta_i^*|y,\psi_{i-1}^*,\psi^{i+1})q(\theta_i,\theta_i^*|y,\psi_{i-1}^*,\psi^{i+1})\}}{E_2\{\alpha_{MH}(\theta_i^*,\theta_i|y,\psi_{i-1}^*,\psi^{i+1})\}}, \tag{3}$$

where $E_1$ is the expectation under the conditional posterior $\pi(\psi^i|y, \psi^*_{i-1})$ and $E_2$ is that under the conditional product measure $\pi(\psi^{i+1}|y, \psi^*_i)q(\theta^*_i, \theta_i|y, \psi^*_{i-1}, \psi^{i+1})$. Here $q(\theta, \theta'|y)$ denotes the candidate generating density of the M–H chain for moving from the current value $\theta$ to a proposed value $\theta'$, and the corresponding M–H probability of accepting the move, $\alpha_{MH}(\theta_i, \theta'_i|y, \psi^*_{i-1}, \psi^{i+1})$, is given by

$$\min\left\{1, \frac{f(y|\theta', \psi^*_{i-1}, \psi^{i+1})\pi(\theta', \psi^*_{i-1}, \psi^{i+1})q(\theta', \theta|y, \psi^*_{i-1}, \psi^{i+1})}{f(y|\theta, \psi^*_{i-1}, \psi^{i+1})\pi(\theta, \psi^*_{i-1}, \psi^{i+1})q(\theta, \theta'|y, \psi^*_{i-1}, \psi^{i+1})}\right\}.$$

Although (3) is a widely applicable formula that can be used in most M–H samplers, it does require knowledge of the normalizing constant of the proposal density $q$. This condition, however, is not satisfied in the ARMH algorithm.

## 3 The ARMH algorithm

Let $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$ denote the target density and let $h(\theta|y)$ denote a source density. In the classical accept–reject method a key requirement is that there exists a constant $c$ such that the condition

$$\mathcal{D} = \{\theta : f(y|\theta)\pi(\theta) \leq ch(\theta|y)\}$$

holds for all $\theta$ in the support $\Theta$ of the target density. The ARMH algorithm is an MCMC sampling procedure in which the domination condition $f(y|\theta)\pi(\theta) \leq ch(\theta|y)$ is not satisfied for some $\theta \in \Theta$, and hence $h(\theta|y)$ is often called a pseudo-dominating density. In this case, let $\mathcal{D}^c$ be the complement of $\mathcal{D}$, and suppose that the current state of the chain is $\theta$. Then the ARMH algorithm is defined as follows.

```
Algorithm 1 One block accept–reject Metropolis–Hastings (ARMH) algorithm
1. A–R step: Generate a draw $\theta' \sim h(\theta|y)$; accept $\theta'$ with probability
       $$\alpha_{AR}(\theta'|y) = \min\left\{1, \frac{f(y|\theta')\pi(\theta')}{ch(\theta'|y)}\right\}.$$
   Continue the process until a draw $\theta'$ has been accepted.
2. M–H step: Given the current value $\theta$ and the proposal value $\theta'$:
   (a) if $\theta \in \mathcal{D}$, set $\alpha_{MH}(\theta, \theta'|y) = 1$;
   (b) if $\theta \in \mathcal{D}^c$ and $\theta' \in \mathcal{D}$, set $\alpha_{MH}(\theta, \theta'|y) = \frac{ch(\theta|y)}{f(y|\theta)\pi(\theta)}$;
   (c) if $\theta \in \mathcal{D}^c$ and $\theta' \in \mathcal{D}^c$, set $\alpha_{MH}(\theta, \theta'|y) = \min\left\{1, \frac{f(y|\theta')\pi(\theta')h(\theta|y)}{f(y|\theta)\pi(\theta)h(\theta'|y)}\right\}$.
   Return $\theta'$ with probability $\alpha_{MH}(\theta, \theta'|y)$. Otherwise return $\theta$.
```

As discussed by Chib and Greenberg (1995), the ARMH algorithm is reversible and, under appropriate regularity conditions, produces draws from the correct density $\pi(\theta|y)$ as the sampling process is iterated. Chib and Greenberg

(1995) also show that the draws produced at the completion of the A–R step have a density

$$q(\theta|y) = d^{-1}\alpha_{AR}(\theta|y)h(\theta|y), \tag{4}$$

which serves as the proposal density for the M–H step. Here $d \equiv \int \alpha_{AR}(\theta|y)$ $h(\theta|y)\,\mathrm{d}\theta$, the normalizing constant of that density, is not available analytically, in contrast to the standard M–H algorithm where the proposal density is fully known. In addition, while a general M–H chain is based on a proposal density $q(\theta, \theta'|y)$, which may depend on the current state $\theta$ as in random walk chains, the ARMH algorithm is an independence chain sampler since the proposal density is of the type $q(\theta'|y) = q(\theta, \theta'|y)$, meaning that it is independent of the current state of the Markov chain. We exploit this feature to simplify our estimation approach.

## 4 Proposed approach

### 4.1 Single-block case

In the single-block case of a general M–H sampler, Chib and Jeliazkov (2001) use the reversibility of the Markov chain to obtain the following expression

$$\pi(\theta^*|y) = \frac{\int \alpha_{MH}(\theta, \theta^*|y)q(\theta, \theta^*|y)\pi(\theta|y)\mathrm{d}\theta}{\int \alpha_{MH}(\theta^*, \theta|y)q(\theta^*, \theta|y)\mathrm{d}\theta}. \tag{5}$$

A simulation consistent estimate of (5) is obtained by averaging $\alpha_{MH}(\theta, \theta^*|y)$ $q(\theta, \theta^*|y)$ in the numerator with draws $\theta \sim \pi(\theta|y)$, while a reduced run provides the draws $\theta \sim q(\theta^*, \theta|y)$ to average $\alpha_{MH}(\theta^*, \theta|y)$ in the denominator. The marginal likelihood estimate can subsequently be calculated by (1). To apply this estimator to the current setting, we substitute (4) into (5), obtaining

$$\pi(\theta^*|y) = \frac{\int \alpha_{MH}(\theta, \theta^*|y)d^{-1}\alpha_{AR}(\theta^*|y)h(\theta^*|y)\pi(\theta|y)\mathrm{d}\theta}{\int \alpha_{MH}(\theta^*, \theta|y)q(\theta|y)\mathrm{d}\theta}. \tag{6}$$

An important simplification of (6) results by letting $\theta^* \in \mathcal{D}$, so that $\alpha_{MH}(\theta^*, \theta|y) = 1$ and $f(y|\theta^*)\pi(\theta^*) \leq ch(\theta^*|y)$. It then follows that (6) can be written as

$$\begin{aligned}\pi(\theta^*|y) &= \frac{f(y|\theta^*)\pi(\theta^*)\int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\mathrm{d}\theta}{cd}\\ &= \frac{f(y|\theta^*)\pi(\theta^*)\int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\mathrm{d}\theta}{c\int \alpha_{AR}(\theta|y)h(\theta|y)\mathrm{d}\theta},\end{aligned}$$

which, upon substitution into (1), produces our first main result that

$$m(y) = \frac{c\int \alpha_{AR}(\theta|y)h(\theta|y)\mathrm{d}\theta}{\int \alpha_{MH}(\theta, \theta^*|y)\pi(\theta|y)\mathrm{d}\theta}. \tag{7}$$

A simulation consistent estimate of $m(y)$, based on (7), can now be obtained as

$$\hat{m}(y) = c\frac{J^{-1}\sum_{j=1}^{J}\alpha_{AR}(\theta^{(j)}|y)}{G^{-1}\sum_{g=1}^{G}\alpha_{MH}(\theta^{(g)},\theta^{*}|y)}, \tag{8}$$

where in the numerator $\theta^{(j)} \sim h(\theta|y)$, and in the denominator $\theta^{(g)} \sim \pi(\theta|y)$. This estimate is particularly simple and uses only quantities which are computed in the course of the ARMH sampling. Therefore, the additional coding and computation for the estimation of the marginal likelihood are minimal.

We make several additional remarks. First, the two quantities in (8) come from the same MCMC run – the draws from $\pi(\theta|y)$ are obtained by accepting or rejecting the candidates from $h(\theta|y)$. Thus the fact that the ARMH algorithm is an independence chain M–H algorithm eliminates the need for a reduced run in the estimation of the marginal likelihood. Second, because the draws from $h(\theta|y)$ are independent and identically distributed (iid), while those from $\pi(\theta|y)$ are generally closer to iid under ARMH sampling than under independence chain M–H sampling, the variance of the resulting estimate will generally be lower. Third, despite its simplicity, this estimator can be applied to many Bayesian models, because the ARMH algorithm does not require that conjugacy be maintained in order to sample from the posterior.

### 4.2 Multi-block case

Grouping all parameters into one block is often a good strategy, but if the dimensionality of the parameter space is large, or if one wishes to exploit the conditional structure of the model to allow for direct sampling, it may be necessary to split the parameter vector into several smaller and more manageable pieces. The current approach readily generalizes to the multi-block case. In fact, the single-block approach is applicable if there are multiple blocks of parameters but only one is sampled by ARMH, since that density ordinate may be estimated at the end of (2), when all other blocks are fixed. Hence, the interesting case is one in which the ARMH output is used to estimate one or more of the reduced conditional density ordinates in (2).

Under the notation introduced in Section 2, let the A–R proposal density be $h(\theta_i|y, \psi_{i-1}, \psi^{i+1})$, which is allowed to depend on the data and the remaining parameters. Now, in the sampling of the $i$th block $\theta_i$, the region of domination is

$$\mathcal{D}_i = \left\{\theta_i : f\left(y|\psi_{i-1},\psi^{i+1}\right)\pi(\theta_i|\psi_{i-1},\psi^{i+1}) \leq c_i\left(\psi_{i-1},\psi^{i+1}\right)h\left(\theta_i|y,\psi_{i-1},\psi^{i+1}\right)\right\},$$

which is generally block and iteration-specific. The M–H proposal density in the $i$th reduced run takes the form

$$q\left(\theta_i|y,\psi^{*}_{i-1},\psi^{i+1}\right) = \frac{\alpha_{AR}\left(\theta^{*}_i|y,\psi^{*}_{i-1},\psi^{i+1}\right)h\left(\theta^{*}_i|y,\psi^{*}_{i-1},\psi^{i+1}\right)}{d\left(y,\psi^{*}_{i-1},\psi^{i+1}\right)},$$

where $d\left(y,\psi^{*}_{i-1},\psi^{i+1}\right)$ is the unknown normalizing constant of $q\left(\theta_i|y,\psi^{*}_{i-1},\psi^{i+1}\right)$. It can easily be shown that the Markov chain reversibility condition used by Chib and Jeliazkov (2001) to obtain (3) can be re-written in terms of $q\left(\theta_i|y,\psi^{*}_{i-1},\psi^{i+1}\right)$, and that its normalizing constant $d\big(y, \psi_{i-1}^{*}, \psi^{i+1}\big)$, being the same on both sides of the reversibility equation, will cancel, so that upon integration $\pi(\theta_i^*|y, \psi_{i-1}^*)$ equals

$$\frac{E_1\big\{\alpha_{MH}\big(\theta_i, \theta_i^*|y, \psi_{i-1}^*, \psi^{i+1}\big)\alpha_{AR}\big(\theta_i^*|y, \psi_{i-1}^*, \psi^{i+1}\big)h\big(\theta_i^*|y, \psi_{i-1}^*, \psi^{i+1}\big)\big\}}{E_2\big\{\alpha_{MH}\big(\theta_i^*, \theta_i|y, \psi_{i-1}^*, \psi^{i+1}\big)\alpha_{AR}\big(\theta_i|y, \psi_{i-1}^*, \psi^{i+1}\big)\big\}},$$

where $E_1$ is the expectation with respect to the conditional posterior $\pi\big(\psi^i|y, \psi_{i-1}^*\big)$ and $E_2$ is that with respect to the product measure $\pi\big(\psi^{i+1}|y, \psi_i^*\big)h\big(\theta_i|y, \psi_{i-1}^*, \psi^{i+1}\big)$. Because of the changing conditioning sets, in the $i$th block $c_i$ and $\mathcal{D}_i$ are iteration specific, so $\theta_i^*$ will not necessarily be in $\mathcal{D}_i$ for every iteration. We caution against attempting the further simplifications used in the single block case because of the risk of decreasing the efficiency of the sampler, which will occur if domination at $\theta_i^*$ is enforced by an excessively large choice of $c_i$. In summary, we present the method in the following steps.

*Step 1:* Set $\psi_{i-1} = \psi_{i-1}^*$ and sample the set of full conditional distributions $\pi(\theta_k|y, \theta_{-k})$, $k = i,\ldots,B$. Let the generated draws be $\{\theta_i^{(g)}, \ldots, \theta_B^{(g)}\}$, $g = 1,\ldots,G$.

*Step 2:* Fix $\theta_i$ at $\theta_i^*$ in the conditioning set to produce $\psi_i^* = (\psi_{i-1}^*, \theta_i^*)$, and sample the remaining distributions $\pi(\theta_k|y, \theta_{-k})$, $k = i + 1,\ldots,B$, to generate $\{\theta_{i+1}^{(j)}, \ldots, \theta_B^{(j)}\}$, $j = 1,\ldots,G$. At each step of the sampling also draw $\theta_i^{(j)} \sim h\Big(\theta_i|y, \psi_{i-1}^*, \psi^{i+1,(j)}\Big)$.

*Step 3:* Estimate the reduced ordinate $\hat{\pi}(\theta_i^*|y, \psi_{i-1}^*)$ as

$$\frac{\frac{1}{G}\sum_{g=1}^{G}\alpha_{MH}(\theta_i^{(g)}, \theta_i^*|y, \psi_{i-1}^*, \psi^{i+1,(g)})\alpha_{AR}(\theta_i^*|y, \psi_{i-1}^*, \psi^{i+1,(g)})h(\theta_i^*|y, \psi_{i-1}^*, \psi^{i+1,(g)})}{\frac{1}{G}\sum_{j=1}^{G}\alpha_{MH}(\theta_i^*, \theta_i^{(j)}|y, \psi_{i-1}^*, \psi^{i+1,(j)})\alpha_{AR}(\theta_i^{(j)}|y, \psi_{i-1}^*, \psi^{i+1,(j)})}. \tag{9}$$

*Step 4:* Estimate the marginal likelihood on the log scale as $\log\hat{m}(y) = \log f(y|\theta^*) + \log\pi(\theta^*) - \sum_{i=1}^{B}\log\hat{\pi}(\theta_i^*|y, \theta_1^*, \ldots, \theta_{i-1}^*)$.

Therefore, in the multi-block ARMH setting, the marginal likelihood estimate is readily available after a straightforward modification of the technique in Chib and Jeliazkov (2001). The approach is also easily applicable in conjunction with other MCMC algorithms, such as M–H or direct sampling from the full-conditionals.

### *4.3 Numerical standard error of the marginal likelihood estimate*

In this section we discuss how the numerical standard error (nse) of the marginal likelihood estimate can be derived. The nse gives the variation that can be expected in the marginal likelihood estimate if the simulation were to be repeated. We specifically focus on the calculation of the nse for the one-block case of Section 4.1, and show that the multi-block case can be handled by existing methods.

There are two complications in estimating the variance of the ratio in (8). One obvious problem is that the lengths of the series of draws from the pseudo-dominating and target densities are different, and hence one can not directly compute the covariance between the numerator and denominator draws. Second, in considering the variability of an estimate obtained by (8), one has to account for the variability in the numerator sample size J. We deal with these problems by applying an approach based on the method of batch means. The denominator quantities $\alpha_{MH}(\theta_1^{(g)}, \theta_1^*|y)$, $g = 1,\ldots,G$, are batched, or sectioned, into $v$ non-overlapping subsamples of length $m$ with $v = G/m$. Each of the denominator subsamples is matched with the draws from the A–R step that were necessary to produce it, thus forming the corresponding $v$ non-overlapping numerator batches of length $n_i \geq m$, $i = 1,\ldots,v$, with $\sum_{i=1}^{v} n_i = J$. Denote the batch means of the numerator quantities by $\{N_i\}$, and those in the denominator by $\{D_i\}$, and let $B_i = N_i/D_i$, $i = 1,\ldots,v$. Then the variance of

$$a = \left\{ \frac{J^{-1}\sum_{j=1}^{J} \alpha_{AR}(\theta^{(j)}|y)}{G^{-1}\sum_{g=1}^{G} \alpha_{MH}(\theta^{(g)}, \theta^*|y)} \right\}$$

is estimated as $\mathrm{var}(a) = \mathrm{var}(B_i)/v$. The batch length $m$ is chosen large enough to guarantee that the first order serial correlation between the batch means is less than 0.05, and to avoid small values of $D_i$ that may produce explosive $B_i$ (in the two examples below, we used $m = 250$). The variance of the log marginal likelihood can be found by the delta method as $\mathrm{var}(\log \hat{m}(y)) = \mathrm{var}(a)/a^2$. The nse of the log marginal likelihood estimate is the square root of $\mathrm{var}(\log \hat{m}(y))$.

Extending the calculation of the nse to the multi-block setting is straightforward by following Chib (1995) for blocks which are sampled directly, and by following Chib and Jeliazkov (2001) for blocks sampled by the M–H algorithm. We emphasize that the latter approach is also applicable to the multi-block ARMH case of Section 4.2, since the numerator and denominator series in (9) have equal lengths. If the prior or likelihood ordinates need to be estimated, then the variance of these estimates must be incorporated by a separate calculation.

We mention that the accuracy of the proposed approach for estimating the nse has been verified in the subsequent examples by repeating the posterior simulations 100 times. The variability of the marginal likelihood estimates from the replications closely mirrored those from the above approach, thus providing a useful validation of this method.

## 5 Examples

We apply the above methods in the context of a multinomial logit and a Gaussian regression model, and illustrate the impact of several ARMH designs on the performance of the MCMC sampler and the marginal likelihood estimation approach. The modelling employs non-conjugate priors because a researcher may wish to incorporate prior information in a more flexible way than that afforded by some particular family of conjugate distributions (as in the Gaussian case) or because such conjugate priors may simply be unavailable (as in the logit model). With non-conjugate priors, however, there is no guarantee that the posterior or its full conditionals will be well behaved, e.g. they could exhibit multimodality, skewness, or kurtosis (O'Hagan, 1994, Chapter 3), thus complicating estimation and marginal likelihood computation. Fortunately, ARMH sampling is well suited to these settings as it does not require conjugate priors or global domination of the proposal over the posterior; the ARMH algorithm also tends to produce MCMC draws that are closer to iid than those from a similarly constructed M–H chain.

### *5.1 Data and models*

Our first example deals with a discrete choice model, the multinomial logit, which has been widely used in many fields of economics (Train, 2003). Specifically, we consider estimation and marginal likelihood computation in the context studied by Small (1982) and Brownstone and Small (1989), where 522 San Francisco Bay Area commuters reported a regular time of arrival relative to the official work start time. These arrival times (ranging between 42.5 minutes early and 17.5 minutes late) are grouped into twelve 5-minute intervals and the probability that commuter *i*'s arrival interval is *t* is modelled as $p_{it} = e^{x_{it}'\beta} / \sum_{k=1}^{12} e^{x_{ik}'\beta}$, where the characteristics $x$ include 13 socioeconomic, behavioral, and transportation-mode variables as described in Appendix 1. Under the prior $\beta \sim N_{13}(\beta_0, B_0)$, the posterior density is given by $\pi(\beta|y) \propto f_N(\beta|\beta_0, B_0) \prod_{i=1}^{522} \prod_{t=1}^{12} p_{it}^{y_{it}}$, where $y_{it} \in \{0, 1\}$ is the binary variable indicating whether individual $i$ chose alternative $t$ and $f_N(\cdot|\cdot)$ denotes the normal density. The posterior can not be sampled directly as it does not belong to a known family of distributions.

Our second application is based on a simple wage determination model using data from Mroz (1987). The goal is to estimate a wage offer function for a sample of 428 employed married women, conditional on four covariates – an intercept, experience in the labor market, experience squared, and education. Linear Gaussian models have been well studied under the usual (normal-gamma) conjugate and semi-conjugate priors (Zellner, 1971). Here, however, we allow for an added degree of flexibility in assessing the prior information and discuss estimation under heavy-tailed non-conjugate prior distributions. Specifically, for $i = 1,\ldots,428$, the model is given by $y_i = x_i'\beta + \varepsilon_i$, where $y_i$ represents woman *i*'s log-wage, $x_i$ is her vector of covariates, and $\varepsilon_i \sim N(0, \sigma^2)$. The priors for the parameters $\beta$ and $\sigma^2$ are given by the multivariate-*t* and the log-*t* distributions $\beta \sim T_{v_b}(\beta_0, B_0)$ and $\sigma^2 \sim \log T_{v_s}(s_0, S_0)$, respectively. These priors allow for additional flexibility by varying the tail behavior through the degrees of freedom parameters $v_b$ and $v_s$.

To illustrate this point, Figure 1 shows the log-pdfs of the inverse gamma, the lognormal, and two log-*t* densities with 5 and 40 degrees of freedom (the means and variances match those of the lognormal distribution). The figure shows that the inverse gamma assigns less mass in the left tail of the distribution than any of the other alternatives; depending on the degrees of freedom parameter, the log-*t* density

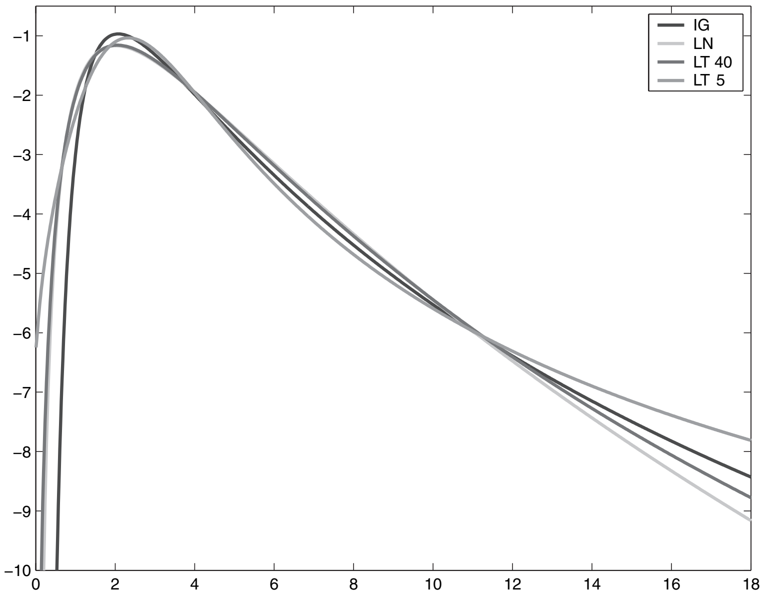

Fig. 1. Log-pdfs for the inverse gamma, lognormal, and two log-$t$ densities.

can have a right tail that is either heavier or thinner than that of the inverse gamma. As expected, the log-$t$ density approaches the lognormal as $v_s$ becomes large. Other general priors are also conceivable and can be handled similarly.

### *5.2 Implementation*

In many models estimated by ARMH, including the two discussed above, it is possible to sample the posterior distribution $\pi(\theta|y)$ in one block by using a tailored source density $h(\theta|y) = f_T(\theta|\mu, \tau V, v)$, where $f_T(\cdot|\cdot)$ denotes a multivariate-$t$ density with mean $\mu$, symmetric positive definite scale matrix $\tau V$ (with $\tau$ being a tuning parameter whose role is illustrated below), and $v$ degrees of freedom. We take $\mu$ as the mode of the log-likelihood and $V$ to be the inverse of the negative Hessian of the log-likelihood evaluated at $\mu$, and set $v = 10$. Having obtained the proposal density, posterior simulation is conducted as in Section 3, while the marginal likelihood and its nse are obtained as in Sections 4.1 and 4.3, respectively.

Often, however, one may also wish to exploit the conditional structure of the model and use the multi-block algorithm of Section 4.2. In the Gaussian model, the multi-block approach is applied, quite naturally, by treating $\beta$ and $\sigma^2$ as separate blocks and sampling proceeds by iteratively drawing from $\pi(\beta|y, \sigma^2)$ and $\pi(\sigma^2|y, \beta)$. The conditional pseudo-dominating densities are taken to be $h(\beta|y, \sigma^2) = f_T(\beta|\mu_1, \tau_1 V_1, v_1)$ and $h(\log(\sigma^2)|y, \beta) = f_T(\log(\sigma^2)|\mu_2, \tau_2 V_2, v_2)$, where we take the parameters $\mu_1$, $V_1$, $\mu_2$, and $V_2$ to be the (analytically available) mode and modal dispersion of the full conditional Bayes updates under non-informative priors. The marginal likelihood is then estimated by writing (1) as

$$m(y) = \frac{\pi(\beta^*)}{\pi(\beta^*|y)} \times \frac{f(y|\beta^*, \sigma^{2*})\pi(\sigma^{2*}|\beta^*)}{\pi(\sigma^{2*}|y, \beta^*)}, \tag{10}$$

where an estimate $\pi(\beta^*|y)$ can be obtained by (9), and the second fraction on the right-hand side can be estimated as

$$c_2(\beta^*)\left\{\frac{J^{-1}\sum_{j=1}^{J}\alpha_{AR}(\sigma^{2(j)}|y, \beta^*)}{G^{-1}\sum_{g=1}^{G}\alpha_{MH}(\sigma^{2(g)}, \sigma^{2*}|y, \beta^*)}\right\},$$

which is a simple application of the one-block estimator (8), since $\beta$ is fixed at $\beta^*$.

In the logit model, Appendix 1 suggests certain natural groupings, based on covariate types, that can be used to partition the 13 × 1 vector $\beta$. To construct a two-block algorithm, we collect the coefficients on the reporting error and travel time variables in $\beta_1(5 \times 1)$ and those on the early and late arrival covariates in $\beta_2(8 \times 1)$. We use the conditional pseudo-dominating densities $h(\beta_1|y, \beta_2) = f_T(\beta_1|\mu_{1|2}, \tau V_{1|2}, v)$ and $h(\beta_2|y, \beta_1) = f_T(\beta_2|\mu_{2|1}, \tau V_{2|1}, v)$ where the parameters $\mu_{1|2}$, $V_{1|2}$, $\mu_{2|1}$, and $V_{2|1}$ are obtained from the overall mode $\mu$ and modal dispersion matrix $V$ of the single block case, using as a rough approximation the conditional updates for the moments of a Gaussian distribution. This method of tailoring performed competitively in our example relative to tailoring by optimization at each iteration, and is considerably faster and less demanding. Hence, $m(y)$ is estimated similarly to (10), using (9) to compute $\pi(\beta_1^*|y)$ and (8) to estimate $f(y|\beta_1^*, \beta_2^*)\pi(\beta_2^*|\beta_1^*)/\pi(\beta_2^*|y, \beta_1^*)$.

The performance of the above algorithm designs can be illustrated by the inefficiency factors for the sampled parameters. The inefficiency factors are calculated as $1 + 2\sum_{l=1}^{L}\rho_k(l)$, where $\rho_k(l)$ is the sample autocorrelation for the $k$th parameter at lag $l$, and $L$ is chosen at values where the autocorrelations taper off. The inefficiency factors approximate the ratio of the numerical variance of the posterior mean from the MCMC chain relative to that from hypothetical iid draws. We consider three one-block ARMH Markov chains with different degrees of pseudo-domination, using the quantity $p = (ch)/(f\pi)$ to represent the relative heights at $\mu$, which, together with the tuning parameter $\tau$ determine the region of domination. Larger values of $p$ and $\tau$ produce larger regions of domination. In Figure 2, we illustrate the inefficiency factors for three settings of the tuning parameters, namely $(\tau = 1, p = 1.25)$, $(\tau = 1.5, p = 1.5)$, and $(\tau = 2, p = 1.75)$. The figure shows that ARMH simulation is generally efficient and the sample becomes essentially iid as $\tau$ and $p$ are increased.

We emphasize that the choice of blocking is not unique and is something that should be determined by the researcher in the context of the particular model and data under consideration. Figure 2 indeed shows that one-block sampling can produce more efficient samples in some settings, but that multi-block simulation can dominate in others. In practice it is useful to (*i*) group parameters that are correlated into one block and sample them jointly, and (*ii*) group parameters in a way that allows for easy construction of suitable pseudo-dominating densities.

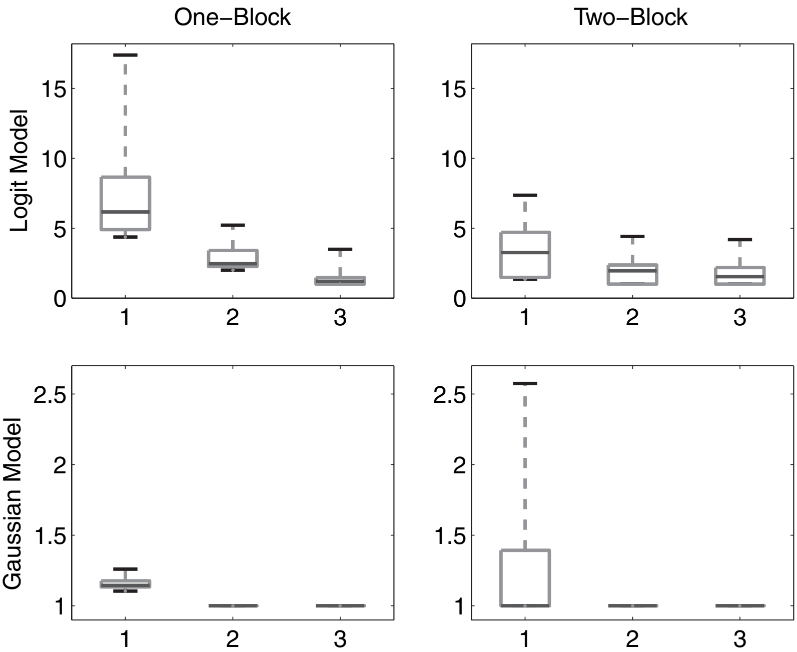

Fig. 2. Inefficiency factors of the one- and two-block logit and Gaussian model parameters for three settings of the tuning parameters $\tau$ and $p$.

It is interesting to look at the posterior distributions of the logit coefficients $\beta_9$ and $\beta_{10}$ (on the late arrival variables $SDL$ and $SDL \cdot WC$, respectively). Both marginal posteriors, shown in Figure 3, are non-Gaussian and skewed, with respective skewness coefficients of –0.59 and 0.59. In cases like these, frequentist asymptotic approximations for constructing confidence intervals (also shown in Figure 3), as well as Bayesian approximations of the marginal likelihood assuming that the posterior is approximately normal (DiCiccio *et al.*, 1997) may be inaccurate. Here,

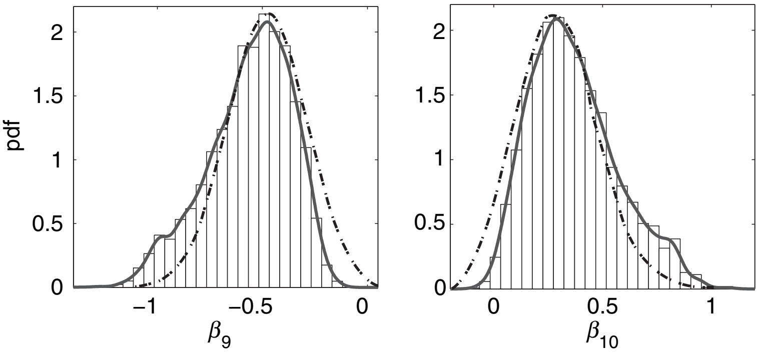

Fig. 3. Two parameters in the logit model: posterior distributions (solid lines) and frequentist asymptotic distributions (dotted lines).

Table 1. Log marginal likelihood estimates for the Gaussian and logit models.

| | Simulation designs | | |
|---|---|---|---|
| | ($\tau = 1, p = 1.25$) | ($\tau = 1.5, p = 1.5$) | ($\tau = 2, p = 1.75$) |
| One-block logit model | | | |
| $\log \hat{m}(y)$ | −1017.180 | −1017.233 | −1017.220 |
| nse | (0.033) | (0.012) | (0.007) |
| Two-block logit model | | | |
| $\log \hat{m}(y)$ | −1017.212 | −1017.237 | −1017.207 |
| nse | (0.027) | (0.009) | (0.011) |
| One-block Gaussian model | | | |
| $\log \hat{m}(y)$ | −458.590 | −458.576 | −458.584 |
| nse | (0.003) | (0.003) | (0.004) |
| Two-block Gaussian model | | | |
| $\log \hat{m}(y)$ | −458.582 | −458.581 | −458.587 |
| nse | (0.006) | (0.006) | (0.006) |

the frequentist estimates would tend to understate blue-collar commuters' desire to avoid arriving too late; they also imply an understated and statistically insignificant (at the 10% level of significance for a two-sided test) difference between the impact of *SDL* on blue-collar and white-collar commuters. In contrast, the posterior distribution assigns probability of 0.994 to positive $\beta_{10}$.

In Table 1, we present the log marginal likelihood and nse estimates corresponding to the simulations from Figure 2. We see that the variability of the estimates is very small and decreases when the regions of domination increase. But the computational loads increase as well for large $\tau$ and $p$ – approximately 11 000 A–R draws are needed to generate the 10 000 ARMH draws for the smallest values of $\tau$ and $p$, about 1.5 times as many are needed for the intermediate values, but about five times that number is needed for the largest $\tau$ and $p$. This illustrates the trade-off between numerical and statistical efficiency that is inherent in the ARMH sampler – as the region of domination becomes larger more draws in the A–R step are needed to generate a given sample, but that sample tends to be closer to iid, thus producing more efficient parameter and marginal likelihood estimates than a typical M–H algorithm.

## 6 Discussion

This paper has presented a method for estimating the marginal likelihood from the building blocks of the ARMH algorithm. The approach is based on the general framework of Chib (1995) and the extension considered in Chib and Jeliazkov (2001), where some of the full-conditional densities are intractable and simulation requires the M–H algorithm. The current method deals with the presence of an unknown normalizing constant in the proposal density and overcomes the difficulties in determining the variability of the marginal likelihood estimate. We show that this estimate and its variability are straightforward to obtain from the output of the ARMH sampler. In two examples, we discuss implementation issues under several ARMH designs and show that the techniques are efficient and generally applicable.

## Appendix 1: Covariates in the largest model in Small (1982)

The analysis uses 13 covariates of four types – reporting error ($R10$ and $R15$), travel time ($TIM$, $TIM{\cdot}SGL$, and $TIM{\cdot}CP$), early arrival ($SDE$, $SDE{\cdot}SGL$, and $SDE{\cdot}CP$), and late arrival ($SDL$, $SDL{\cdot}WC$, $SDLX$, $D1L{\cdot}WC$, and $D2L$). In the preceding, $SD$ is Schedule Delay, i.e. arrival time minus official work start time rounded to nearest 5 minutes ($SD = \{-40, -35, \ldots, 10, 15\}$); $R10 = 1\{SD = -40, -30, -20, -10, 0, 10\}$; $R15 = 1\{SD = -30, -15, 0, 15\}$; $TIM$ is Travel time in minutes; $SDE = \max\{-SD, 0\}$; $SDL = \max\{SD, 0\}$; $D1L = 1\{SD \geq 0\}$; $FLEX$ is reported flexibility for arriving time; $D2L = 1\{SD \geq FLEX\}$; $SGL$ is a dummy for a one-person household; $CP$ is carpool dummy reconstructed in Brownstone and Small (1989) to account for previously missing data; $WC$ is a dummy for a white collar worker; and $SDLX = \max\{SD - FLEX, 0\}$. For further details and some alternative models, see Small (1982) and Brownstone and Small (1989).

**Acknowledgements**

We thank Professors Edward Greenberg, Itai Sened and Kenneth Small for valuable comments on earlier drafts. Financial assistance from the Department of Economics at Washington University is gratefully acknowledged.

## References

Basu, S. and S. Chib (2003), Marginal likelihood and Bayes Factors for Dirichlet process mixture models, *Journal of the American Statistical Association* **98**, 224–235.

Brownstone, D. and K. Small (1989), Efficient estimation of nested logit models, *Journal of Business & Economic Statistics* **7**, 67–74.

Carlin, B. P. and S. Chib (1995), Bayesian model choice via Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society* B **57**, 473–484.

Chen, M. H. and Q. M. Shao (1997), On Monte Carlo methods for estimating ratios of normalizing constants, *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 E. Greenberg (1995), Understanding the Metropolis–Hastings algorithm, *American Statistician* **49**, 327–335.

Chib, S. and I. Jeliazkov (2001), Marginal likelihood from the Metropolis–Hastings output, *Journal of the American Statistical Association* **96**, 270–281.

DiCiccio, T. J., R. E. Kass, A. E. Raftery and L. Wasserman (1997), Computing Bayes factors by combining simulation and asymptotic approximations, *Journal of the American Statistical Association* **92**, 903–915.

Gelfand, A. E. and D. Dey (1994), Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society* B **56**, 501–514.

Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

Han, C. and B. P. Carlin (2001), Markov chain Monte Carlo methods for computing Bayes factors: a comparative review, *Journal of the American Statistical Association* **96**, 1122–1132.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

Jeffreys, H. (1961), *Theory of probability* (3rd edn), Clarendon Press, Oxford.

Meng, X. L. and W. H. Wong (1996), Simulating ratios of normalizing constants via a simple identity: a theoretical exploration, *Statistica Sinica* **6**, 831–860.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

Mroz, T. (1987), The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica* **55**, 765–799.

Newton, M. A. and A. E. Raftery (1994), Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society* B **56**, 3–48.

O'Hagan, A. (1994), *Kendall's advanced theory of statistics, Vol. 2B*, John Wiley & Sons, New York.

Small, K. (1982), The scheduling of consumer activities: work trips, *The American Economic Review* **72**, 3, 467–479.

Tierney, L. (1994), Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1762.

Train, K. (2003), *Discrete choice methods with simulation*, Cambridge University Press, Cambridge, UK.

Verdinelli, I. and L. Wasserman (1995), Computing Bayes factors using a generalization of the Savage–Dickey density ratio, *Journal of the American Statistical Association* **90**, 614–618.

Zellner, A. (1971), *An introduction to Bayesian inference in econometrics*, John Wiley & Sons, New York.

Received: June 2004. Revised: December 2004.