Model choice with MCMC on product spaces without using pseudo-priors

Peter Green and Tony O'Hagan
University of Bristol and University of Nottingham

March 8, 2000

Abstract

Carlin and Chib proposed a Markov chain Monte Carlo method for exploring alternative models. We show that, contrary to a claim by Carlin and Chib, it is not necessary to sample from the pseudo-priors which are a feature of their approach. Although this may improve the efficiency of the Carlin and Chib method, examination of one of their examples suggests that the reversible jump method of Green is considerably better. For this example, we also calculate the fractional Bayes factor, which avoids the specification of artificial proper priors, which are needed by both the Carlin and Chib method and reversible jump.

1 General framework

A fundamental problem in statistics is the comparison of alternative models, often for the purpose of choosing a most appropriate model. Within the Bayesian framework the key tool for comparing models is the Bayes factor. Recently, two Bayesian methods based upon Markov chain Monte Carlo methodology have been proposed, by Carlin and Chib (1995) and Green (1995), to compute Bayes factors. We consider first the method of Carlin and Chib, and demonstrate that their device of pseudo-priors is unnecessary. We then compare their method with that of Green, using one of Carlin and Chib's example datasets. Finally, we also exhibit the fractional Bayes factor

of O'Hagan (1995) for this example, and remark on the relative computational costs of the different approaches.

In Carlin and Chib's product-space approach to MCMC computation for Bayesian model choice, artificial "pseudo-priors" have to be specified, both to define their hierarchical model, and for use in updating model parameters. In this note, we show that their approach can be modified to *avoid* using pseudo-priors for this second purpose, while still maintaining the required equilibrium distribution. The consequences are both that considerable computer time can be saved, and that there is no need to conduct pilot runs tuning these pseudo-priors to achieve reasonable statistical efficiency.

As do Carlin and Chib, we address the model choice problem by means of a hierarchical formulation in which there is a model indicator $k \in \mathcal{K}$, a parameter vector θ_k for each k, and data y. We will write $\boldsymbol{\theta}$ for the extended vector with components $(\theta_i, i \in \mathcal{K})$. Carlin and Chib's approach is to operate on the *product space* containing k and all θ_k simultaneously; it is therefore necessary to specify the distribution of each parameter vector *whatever* the value of the model indicator, in order to define fully the joint distribution of unknowns and data. Thus, we assume given the model probabilities $p(k)$ for $k \in \mathcal{K}$, priors $p(\theta_k|i)$ for all $k, i \in \mathcal{K}$, and likelihoods

$$p(y|k, \boldsymbol{\theta}) = p(y|k, \theta_k); \tag{1}$$

this identity, which is quite natural, is equivalent to assuming that, given the model indicator k, y and θ_k are independent of θ_i for $i \neq k$. In terms of these ingredients, the joint distribution of all variables is

$$p(k, \boldsymbol{\theta}, y) = p(k) \prod_{i \in \mathcal{K}} p(\theta_i|k) p(y|k, \boldsymbol{\theta}). \tag{2}$$

In contrast to Carlin and Chib's formulation, there will be no practical objection to defining $p(\theta_i|k)$ independently of k in our approach.

To sample from the posterior distribution

$$p(k, \theta_k|y) \propto p(k, \boldsymbol{\theta}|y) \propto p(k, \boldsymbol{\theta}, y), \tag{3}$$

we propose MCMC moves of two types: an update of k, and an update of the θ_k *only for the current value of the model indicator* k. Each of these moves preserves the target distribution

$$p(k, \boldsymbol{\theta}|y) \propto p(k, \boldsymbol{\theta}, y)$$

individually, as we shall see below, and the moves can therefore be used alternately, completely at random, or by some means of restricted randomisation. Carlin and Chib use only alternate updates.

To update k, any standard method can be used. At least for a small number of candidate models, it will be convenient to use a Gibbs kernel (as do Carlin and Chib), that is to sample from the full conditional

$$p(k|\boldsymbol{\theta}, y) = \frac{p(k)\prod_{i\in\mathcal{K}} p(\theta_i|k)p(y|k,\boldsymbol{\theta})}{\sum_{k'} p(k')\prod_{i\in\mathcal{K}} p(\theta_i|k')p(y|k',\boldsymbol{\theta})}; \tag{4}$$

alternatively, a Hastings kernel would be even easier to use, and remains practical for any number of models.

To update $\boldsymbol{\theta}$ when k is the current value of the model indicator, all θ_i for $i \neq k$ are left unchanged, in contrast to the method of Carlin and Chib, who propose to update all θ_i from their full conditionals, and thus need to simulate from pseudo-priors. The "active" θ_k is updated using any MCMC kernel, $Q_k(\theta_k \to \theta_k')$ say, that satisfies detailed balance with respect to $p(\theta_k|k)p(y|k,\theta_k)$ (or, equivalently, with respect to $p(\theta_k|k,y)$). Note that it is exactly such kernels that would be used in a reversible sampler for computation for model k on its own. A valid example would be the Gibbs kernel, that draws the new value from $p(\theta_k|k,y)$) itself. Whether or not Gibbs is used, the complete kernel for the $\boldsymbol{\theta}$ update can then be written

$$P(\boldsymbol{\theta} \to \boldsymbol{\theta}') = \sum_{i\in\mathcal{K}} I[k=i]Q_i(\theta_i \to \theta_i')\prod_{j\neq i} I[\theta_j = \theta_j'],$$

where $I[\,]$ denotes the indicator function. The detailed balance condition we are assuming on Q_i says that

$$p(\theta_i|i)p(y|i,\theta_i)Q_i(\theta_i \to \theta_i') = p(\theta_i'|i)p(y|i,\theta_i')Q_i(\theta_i' \to \theta_i)$$

for all θ_i, θ_i'. This can be rewritten using (1) as

$$p(\theta_i|i)p(y|i,\boldsymbol{\theta})Q_i(\theta_i \to \theta_i') = p(\theta_i'|i)p(y|i,\boldsymbol{\theta}')Q_i(\theta_i' \to \theta_i).$$

Multiplying by $p(k)\prod_{j\neq i} p(\theta_j|k) \times I[k=i]\prod_{j\neq i} I[\theta_j = \theta_j']$, and summing over all $i \in \mathcal{K}$, we obtain

$$p(k,\boldsymbol{\theta},y)P(\boldsymbol{\theta} \to \boldsymbol{\theta}') = p(k,\boldsymbol{\theta}',y)P(\boldsymbol{\theta}' \to \boldsymbol{\theta}),$$

that is, we have detailed balance for P, as required.

Essentially the same proof shows that if the Q_i satisfy only *global* balance,

$$\int p(\theta_i|i)p(y|i,\theta_i)Q_i(\theta_i \to \theta_i')d\theta_i = p(\theta_i'|i)p(y|i,\theta_i')$$

then this is also inherited by P:

$$\int p(k,\boldsymbol{\theta},y)P(\boldsymbol{\theta} \to \boldsymbol{\theta}')d\boldsymbol{\theta} = p(k,\boldsymbol{\theta}',y).$$

This is therefore sufficient to ensure that the target distribution (3) is left invariant by P, even though P is no longer reversible.

In drawing information from the realised values of $(k,\boldsymbol{\theta})$, note that from (2), both the marginal distribution of k alone and the conditional distribution of θ_k given the model indicator takes the value k, are correct for each $k \in \mathcal{K}$. They are not affected by the form of the pseudo-priors, which only influence the transitions updating k.

An example of the procedure we have described, in which both k and $\boldsymbol{\theta}$ updates use Gibbs kernels, is mentioned by Carlin and Chib in the last paragraph of their section 2 as a "tempting alternative". There it is apparently suggested that the algorithm might not converge to the correct distribution, or at least, that conventional Markov chain convergence theory is insufficient to establish this. As we see from the proof above, this suggestion is unfounded.

In practice, it should be adequate to set $p(\theta_i|k) = p(\theta_i|i)$ for all k; an advantage of this is that these terms then all cancel from (4). In contrast, in the method of Carlin and Chib, it would seem necessary to set $p(\theta_i|k)$ approximately equal to $p(\theta_i|i,y)$ for $k \neq i$ to ensure good mixing.

2 Example

Carlin and Chib present an example in which there are two alternative regressor variables to explain a single dependent variable. For $n = 42$ specimens of radiata pine, the maximum compressive strength parallel to the grain y_i was observed, along with the specimen's density x_i and its density adjusted for resin content z_i. The two competing models seek to explain strength by

regression either on density or on adjusted density. Thus, Model 1 asserts that

$$y_i = a + \beta(x_i - \bar{x}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

whereas Model 2 asserts that

$$y_i = \gamma + \delta(z_i - \bar{z}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \tau^2),$$

for $i = 1, \ldots, n$. Hence $\boldsymbol{\theta}_1 = (\alpha, \beta, \sigma^2)$ and $\boldsymbol{\theta}_2 = (\gamma, \delta, \tau^2)$. Carlin and Chib assume independent normal priors for α, β, γ, δ, σ^2 and τ^2, specified by

$$\begin{array}{rcllll} \alpha & \sim & N(3000,\, 10^6), & \beta \sim N(185,\, 10^4), & \sigma^2 \sim 600^2 \chi_6^{-2}, \\ \gamma & \sim & N(3000,\, 10^6), & \delta \sim N(185,\, 10^4), & \tau^2 \sim 600^2 \chi_6^{-2}. \end{array}$$

The distributions for σ^2 and τ^2 have means and standard deviations equal to 300^2. We denote the prior probability for Model i by $P(M = i) = \pi_i$. Then the posterior is

$$f(M, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y}, \mathbf{z}) \propto \left\{ \begin{array}{l} \pi_1 \, \sigma^{-n+6} \exp(-Q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_1) \text{ if } M = 1\,, \\ \pi_2 \, \tau^{-n+6} \exp(-Q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_2) \text{ if } M = 2\,, \end{array} \right. \tag{5}$$

where

$$\begin{aligned} Q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_1) \;\; = \;\; & \sigma^{-2} \left\{ 600^2 + \sum\nolimits_{i=1}^{n} (y_i - \alpha - \beta(x_i - \bar{x})^2 \right\} \\ & + 10^{-6}(\alpha - 3000)^2 + 10^{-4}(\beta - 185)^2\,. \end{aligned}$$

The full conditional distributions are easily derived. Those of α, β, γ and δ are normal and those of σ^2 and τ^2 are inverse chi-square. The full conditional distribution for M is just proportional to the two lines of (5). In our MCMC runs we set $\pi_1 = 0.9995$, $\pi_2 = 0.0005$, the same values as those used by Carlin and Chib, who then found a posterior probability of 0.3114 for $M = 1$.

We ran a single chain with updating steps alternating between updating the parameters $\boldsymbol{\theta}_M$ for the current model and updating M. For the M update we used a Metropolis step with proposal to switch M, so that if for instance the current model is Model 1 ($M = 1$), we update to $M = 2$ with probability

$$\min\left\{1, (\pi_2/\pi_1)(\sigma^2/\tau^2)^{3+n/2} \exp\left[\left(Q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_1) - Q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_2)\right)/2\right]\right\}\,.$$

The Metropolis step is preferred to a Gibbs step here because it maximises the chance of changing M. Nevertheless we found that the chain mixes very

badly. In a run of 100,000 iterations (i.e. pairs of updates) the proposal to change M was accepted only 6895 times. Furthermore, the distributions of sojourn times in each model are extremely skewed. The 3448 sojourn times in Model 2, for instance, had a mean of 22.65 but a standard deviation of 346. Sojourn times of 11,230 and 15,802 iterations were observed. Essentially, the ergodic estimate of the posterior distribution for M depends on the observed mean sojourn times in the two models. The mean sojourn time in Model 1 was 6.356, and the resulting estimate of the posterior probability that $M = 1$ is $\frac{6.356}{6.356+22.65} = 0.2191$. But because of the high variances of sojourn times this is a poor estimate. Based on the sample variance of 346, in order to estimate the mean sojourn time in Model 2 with a standard deviation of 0.05 (which is the kind of accuracy needed to match the accuracy claimed by Carlin and Chib) we would need to observe $(346/0.05)^2$, or about 5 million sojourns. Figure 1 plots the cumulative relative frequency of Model 1, and demonstrates the lack of convergence.

The reason for the problem is not hard to find. When the chain switches out of Model i, it is more likely to have done so at a value of $\boldsymbol{\theta}_i$ with a relatively low posterior density. For the chain to switch back to Model i it must go back to the same value of $\boldsymbol{\theta}_i$, which therefore typically has a low chance of occurring. The extremely long sojourns are associated with trying to switch back to a parameter value with very low probability. The Carlin and Chib method avoids this problem by updating the parameters of the models that are not current, so that the chain is not always trying to switch back to the same point. Of course, when there are very many models, this efficiency is achieved at a high price, because the parameters of every model are updated every iteration. Their method will not be viable when the number of competing models is large, such as in complex mixture or partition models, or in variable selection problems with many candidate explanatory variables. In such models, it is a large overhead even to retain current values of parameters in all the possible models.

3 Reversible jump method

An alternative is the reversible jump MCMC method of Green (1995). This method is generally significantly more flexible and more efficient than the Carlin and Chib approach. In fact, their example of competing simple re-

gressions provides an elementary illustration of the reversible jump method. In general, the reversible jump approach considers a move from $(M = i, \boldsymbol{\theta}_i)$ to $(M = j, \boldsymbol{\theta}_j)$ where instead of $\boldsymbol{\theta}_j$ being the last visited value of that parameter it is a (possibly random) function of $\boldsymbol{\theta}_i$. The idea is that the value of the $\boldsymbol{\theta}_j$ proposal should somehow represent a natural correspondence to $\boldsymbol{\theta}_i$. In the present example, both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are parameters in identical regression models using rather similar explanatory variables, and the natural choice of correspondence between the parameters when a change of model is proposed is simply $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. We therefore run a chain which alternates between a Gibbs update on the parameters $\boldsymbol{\theta}_i$ of the current model and a Metropolis step proposing a switch from $(M = i, \boldsymbol{\theta}_i)$ to $(M = j, \boldsymbol{\theta}_j = \boldsymbol{\theta}_i)$. The acceptance probability for a step from Model 1 to Model 2 becomes

$$\min\left\{1, (\pi_2/\pi_1)\exp\left[\left(Q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_1) - Q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_1)\right)/2\right]\right\} \ .$$

A single run of 100,000 of these iterations produces vastly better mixing than the chain in the previous section. Figure 2 plots the cumulative relative frequency of Model 1, which now seems to be converging well. Figure 3 shows the relative frequency of Model 1 in successive blocks of 100 iterations, which indicates almost instant convergence and stability. The autocorrelations of this sequence of 100-iteration means are all negligible. Their mean is 0.29063, which is therefore the estimate of the posterior probability of Model 1. The sample variance of these 1000 blocks is 0.0034543, and therefore the estimated standard error of the estimate is 0.00186. This is as good as the accuracy claimed by Carlin and Chib from a total of 250,000 iterations of their method. Note also that their method takes twice as long per iteration, since both sets of parameters must be updated each time, and this effect becomes much more serious in problems with many competing models.

4 Resolution of computational discrepancies

There is, however, a computational discrepancy to be resolved here. Carlin and Chib give an estimate of 0.3114 with s.e. of 0.00166 for the posterior probability of model 1, whereas our reversible jump estimate is 0.29063 with s.e. 0.00186. The difference of 0.0208 is 8 standard deviations from zero (computing the standard deviation of the difference from the quoted s.e.s and independence), so something is wrong with one or other of the computations.

To resolve this, note that the Bayes factor is

$$B_{12}(y) = \frac{q_1(y)}{q_2(y)}$$

where

$$q_i(y) = \int p(y \mid \theta_i)\, p(\theta_i)\, d\theta_i \,.$$

Now both models are simple linear regression models, but this calculation cannot be done analytically because the priors are not conjugate. Nevertheless, it is easy to integrate out the regression parameters from each model, leaving a one-dimensional integration with respect to the error variance parameter to be done numerically. We obtain

$$q_1(y) = k \int_0^\infty \sigma^{-(n+8)} \left(\frac{n}{\sigma^2} + 10^{-6}\right)^{-0.5} \left(\frac{S_{xx}}{\sigma^2} + 10^{-6}\right)^{-0.5} \exp\left(-\frac{1}{2}Q_3\right)\, d\sigma^2$$

where

$$Q_3 = \left(600^2 + S_{yy.x}\right)\sigma^{-2} + \left(10^6 + \frac{\sigma^2}{n}\right)^{-1} (\bar{y} - 3000)^2 + \left(10^6 + \frac{\sigma^2}{S_{xx}}\right)^{-1} \left(\hat{\beta} - 185\right)^2 ,$$

$S_{xx} = \sum (x_i - \bar{x})^2$, etc., $\hat{\beta} = S_{xx}^{-1} S_{xy}$, $S_{yy.x} = S_{yy} - S_{xx}^{-1} S_{xy}^2$ as usual, and k is a constant. A similar formula applies for $q_2(y)$, just replacing the x_is by z_is (and in particular the same constant k appears). These integrals were carried out by numerical integration using Simpson's rule on 1000 equally spaced ordinates. (As few as 100 function evaluations produced essentially identical results, because of the nice smooth behaviour of the integrands.)

We found $\ln B_{12}(y) = -8.489$. Combining this with the prior probability for model 1 of 0.9995, we find the posterior probability to be 0.29135. We believe that this calculation is essentially exact, and certainly correct to three decimal places. This suggests strongly that our reversible jump computation is correct, and that there must have been a flaw in Carlin and Chib's computation.

5 Fractional Bayes factor

In this example, we have used the same prior distributions as Carlin and Chib. Specifically, they used weak, but proper, prior distributions. In the context

of model comparison, O'Hagan (1995) argues that Bayes factors may be highly sensitive to the specification of prior distributions on the parameters of the competing models, when such prior information is weak. He proposes an alternative method known as the fractional Bayes factor. The Bayes factor is undefined if we use improper priors, which is why Carlin and Chib propose proper priors, but this is the underlying cause of the resulting non-robustness. The fractional Bayes factor (FBF) may be used with improper priors, so avoiding an arbitrary and artificial specification of proper priors.

We computed the FBF for this example, using the improper priors $p(\alpha, \beta, \sigma^2) \propto \sigma^{-2}$ for model 1 and $p(\gamma, \delta, \tau^2) \propto \tau^{-2}$ for model 2. It is now possible to do the calculation analytically, with the result that the fractional Bayes factor using training fraction b is

$$B_{12}^b(y) = \left(\frac{S_{yy.x}}{S_{yy.z}}\right)^{-n(1-b)/2} .$$

We find that the ratio $S_{yy.x}/S_{yy.z}$ of residual sums of squares is 1.501. Perhaps the fairest comparison with the Bayes factor using proper priors is obtained by setting $b = 0$, which is possible in this case and yields the factor $\ln B_{12}^0(y) = -8.529$, and produces a posterior probability of 0.283 for model 2. We would generally prefer this sort of calculation rather than assuming arbitrary weak priors. For this simple example where the FBF can be obtained analytically, the contrast in computing effort is dramatic. The FBF is computed essentially instantaneously. The exact Bayes factor, using one-dimensional integrations, is much slower (although on a modern computer it is also almost instantaneous). The MCMC methods are massively slower, although Carlin and Chib's takes several times as long as reversible jump (and apparently gets the wrong answer).

Note. This work was originally carried out in 1995–6. In view of the fact that the reversible jump approach looked likely to be accepted very quickly as the standard technique for model comparison by MCMC, making the Carlin and Chib work of minor interest, it did not seem necessary for us to seek to publish this paper. However, in view of several requests for details of our work, we have prepared it in this limited publication form as a University of Nottingham Research Report.

References

[1] Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *J. Roy. Statist. Soc. B* **57**, 473–484.

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

[3] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. B* **57**, 99–138.