Biometrika (2006), **93,** 1, *pp.* 41–52

Printed in Great Britain

Efficient Bayes factor estimation from the reversible jump output

By FRANCESCO BARTOLUCCI

Department of Economics, Finance and Statistics, University of Perugia, 06123 Perugia, Italy

bart@stat.unipg.it

LUISA SCACCIA

Department of Economic and Financial Institutions, University of Macerata, 62100 Macerata, Italy

scaccia@unimc.it

and ANTONIETTA MIRA

Department of Economics, University of Insubria, 21100 Varese, Italy

antonietta.mira@uninsubria.it

Summary

We propose a class of estimators of the Bayes factor which is based on an extension of the bridge sampling identity of Meng & Wong (1996) and makes use of the output of the reversible jump algorithm of Green (1995). Within this class we give the optimal estimator and also a suboptimal one which may be simply computed on the basis of the acceptance probabilities used within the reversible jump algorithm for jumping between models. The proposed estimators are very easily computed and lead to a substantial gain of efficiency in estimating the Bayes factor over the standard estimator based on the reversible jump output. This is illustrated through a series of Monte Carlo simulations involving a linear and a logistic regression model.

Some key words: Bayesian model choice; Bridge sampling; Marginal likelihood; Markov chain Monte Carlo; Reversible jump.

1. Introduction

In the Bayesian context, the most widespread model choice criterion is based on the Bayes factor (Jeffreys, 1935, 1961; Kass & Raftery, 1995; Lavine & Schervish, 1999), which, for two models, $\mathscr{M}_1$ and $\mathscr{M}_2$ say, may be interpreted as a measure of the evidence provided by the data in favour of $\mathscr{M}_1$ relative to $\mathscr{M}_2$. It is the ratio of the marginal likelihood of $\mathscr{M}_1$ to that of $\mathscr{M}_2$ or, equivalently, the ratio of the posterior to the prior odds of $\mathscr{M}_1$ relative to $\mathscr{M}_2$. Clearly, when there are more than two models, the Bayes factor can be computed for each pair in order to identify the one which is most strongly supported by the data.

The method's applicability has been limited by the fact that exact evaluation of the Bayes factor is seldom feasible. However, many estimation methods are now available in

the Markov chain Monte Carlo literature; for a review see Han & Carlin (2001), Dellaportas et al. (2002) or Green (2003). These are based on the estimation of marginal likelihoods or on the estimation of the posterior probabilities of the available models. Typically, methods belonging to the first class use the output of separate Markov chains run independently for each model, while those in the second class use the output of a single Markov chain on an enlarged state space that explores all models at once.

One of the earliest methods belonging to the first class was proposed by Chib (1995) and further extended by Chib & Jeliazkov (2001). This method is based on the estimation of the marginal likelihood of any available model from the output of a Markov chain Monte Carlo algorithm used to draw samples from the posterior distribution of the model parameters. Recently, a very powerful tool for estimating the ratio of the normalising constants of two distributions, and so also the Bayes factor, has been introduced by Meng & Wong (1996) on the basis of the so-called bridge sampling identity. Other estimation methods in this first class have been proposed by Chen & Shao (1997a, b), Gelman & Meng (1998) and Meng and Schilling (2002).

The best-known method in the second class is based on the reversible jump algorithm of Green (1995), which allows us to sample from the model and parameter space jointly and estimates the posterior probabilities of each model by the relative frequencies of visits. Other methods in this class have been proposed by Carlin & Chib (1995), Godsill (2001) and Dellaportas et al. (2002).

The above estimation methods present advantages and drawbacks. Those in the first class, in particular, are quite efficient but can be impractical if the number of candidate models is very large. Moreover, it is usually complicated to implement them and they require a fair amount of 'bookkeeping'. Instead, the standard estimator based on the reversible jump output, as well as the other methods in the second class, is easily computed and can be used even when the number of competing models is huge; however, this estimator is generally not very efficient. Moreover, the reversible jump algorithm, on which it is based, requires accurate tuning of the jump proposals in order to promote mixing among models.

In this paper, we propose a class of estimators of the Bayes factor based on the reversible jump output, which improves the efficiency of the standard estimator. The approach is based on an extension of the bridge sampling identity of Meng & Wong (1996) along the same lines as Chen & Shao (1997b), Meng & Schilling (2002) and Mira & Nicholls (2004). Within the proposed class of estimators we derive, on the basis of the optimal rule given in Meng & Wong (1996), the most efficient estimator that may be computed through a simple iterative rule. We also suggest a suboptimal estimator which may be directly computed, without extra computing effort, on the basis of the acceptance probabilities used within the reversible jump algorithm for jumping between models.

2. Preliminaries

2·1. *Definition of Bayes factor*

Let $\{\mathscr{M}_1, \ldots, \mathscr{M}_K\}$ denote the set of available models and, for model $\mathscr{M}_k$, let Θ_k be the parameter space, whose elements will be denoted by θ_k with the subscript k dropped when the model to which is referred is clear from the context. Also let $p_k(y|\theta)$ be the likelihood for an observed sample y, let $p_k(\theta)$ be the prior distribution on the parameters and let p_k

be the prior probability of $\mathscr{M}_k$. The Bayes factor between $\mathscr{M}_k$ and $\mathscr{M}_l$ $(l \neq k)$ is defined as

$$B_{kl} = \frac{p_k(y)}{p_l(y)}, \tag{1}$$

where

$$p_k(y) = \int_{\Theta_k} p_k(y|\theta) p_k(\theta) d\theta \tag{2}$$

is the marginal likelihood of $\mathscr{M}_k$. This justifies the common interpretation of the Bayes factor according to which the larger B_{kl} is, the greater is the evidence provided by the data in favour of $\mathscr{M}_k$ relative to $\mathscr{M}_l$. An alternative expression is

$$B_{kl} = \frac{p(k|y)}{p(l|y)} \Big/ \frac{p_k}{p_l}, \tag{3}$$

where

$$p(k|y) = \frac{p_k(y) p_k}{\sum_{h=1}^{K} p_h(y) p_h} \tag{4}$$

is the posterior probability of $\mathscr{M}_k$. Note that (3) is equal to the ratio of the posterior to the prior odds in favour of $\mathscr{M}_k$.

2·2. *Estimating marginal likelihoods*

Two methods that received considerable attention among those dealing with the estimation of marginal likelihoods are due to Chib (1995) and to Meng & Wong (1996). The first is based on the identity

$$p_k(y) = \frac{p_k(y, \bar{\theta})}{p_k(\bar{\theta}|y)},$$

which holds for any fixed $\bar{\theta} \in \Theta_k$, where $p_k(y, \theta) = p_k(y|\theta) p_k(\theta)$ and

$$p_k(\theta|y) = \frac{p_k(y, \theta)}{p_k(y)} \tag{5}$$

is the posterior distribution of the parameters of model $\mathscr{M}_k$. Estimating $p_k(y)$ is therefore equivalent to estimating $p_k(\bar{\theta}|y)$ for an appropriately chosen $\bar{\theta}$. Chib & Jeliazkov (2001) showed that a suitable estimate of $p_k(\bar{\theta}|y)$ may be obtained from the Metropolis–Hastings (Metropolis et al., 1953; Hastings, 1970) output for sampling from the posterior distribution $p_k(\theta|y)$ that uses

$$\alpha_k(\theta, \theta^*) = \min\left\{1, \frac{p_k(y, \theta^*) q_k(\theta|\theta^*)}{p_k(y, \theta) q_k(\theta^*|\theta)}\right\}$$

as acceptance probability for moving from θ to a proposed θ^*, where $q_k(\theta^*|\theta)$ is the proposal distribution. In fact, we have

$$p_k(\bar{\theta}|y) = \frac{E_k\{\alpha_k(\theta, \bar{\theta}) q_k(\bar{\theta}|\theta)\}}{E_k^*\{\alpha_k(\bar{\theta}, \theta^*)\}}, \tag{6}$$

where E_k denotes expectation under $p_k(\theta|y)$ and E_k^* that under $q_k(\theta^*|\bar{\theta})$. Consequently, we can estimate $p_k(\bar{\theta}|y)$ by

$$\hat{p}_k(\bar{\theta}|y) = \frac{\sum_{i=1}^{n_k} \alpha_k(\theta_{ki}, \bar{\theta}) q_k(\bar{\theta}|\theta_{ki})/n_k}{\sum_{i=1}^{n_k^*} \alpha_k(\bar{\theta}, \theta_{ki}^*)/n_k^*}, \tag{7}$$

where θ_{ki} $(i = 1, \ldots, n_k)$ is a sample from $p_k(\theta|y)$ and θ_{ki}^* $(i = 1, \ldots, n_k^*)$ is a sample from $q_k(\theta^*|\bar{\theta})$. Then we can estimate the marginal likelihood of $\mathscr{M}_k$ as $\hat{p}_k(y) = p_k(y, \bar{\theta})/\hat{p}_k(\bar{\theta}|y)$, and, because of (1), the Bayes factor between $\mathscr{M}_k$ and $\mathscr{M}_l$ as

$$\hat{B}_{kl}^{\mathrm{CJ}} = \frac{\hat{p}_k(y)}{\hat{p}_l(y)}.$$

In order to increase the estimator efficiency, Chib & Jeliazkov (2001) also suggested splitting the parameters into blocks which are updated separately. The point $\bar{\theta}$ is arbitrary, but in practice it is chosen as a point of high posterior density so as to maximise the accuracy of the approximation; it generally coincides with the posterior mean or the maximum likelihood estimate of θ_k.

The approach of Meng & Wong (1996), based on the bridge sampling identity, allows us to estimate the ratio of the normalising constants of two distributions. From (5) we have that the marginal likelihood of a model may be seen as the normalising constant of the posterior distribution, and so this approach may be used directly to estimate B_{kl} as follows. Assume for simplicity that Θ_k and Θ_l have the same dimension; then, according to the bridge sampling identity,

$$B_{kl} = \frac{p_k(y)}{p_l(y)} = \frac{E_l\{h(\theta)p_k(y, \theta)\}}{E_k\{h(\theta)p_l(y, \theta)\}}, \tag{8}$$

where $h(\theta)$ is an arbitrary function such that the expected values above are bounded away from 0 and ∞. Therefore, given random draws θ_{ki} $(i = 1, \ldots, n_k)$ from $p_k(\theta|y)$ and θ_{li} $(i = 1, \ldots, n_l)$ from $p_l(\theta|y)$, together with a choice of $h(\theta)$, we can estimate the Bayes factor by

$$\hat{B}_{kl}^{\mathrm{MW}} = \frac{\sum_{i=1}^{n_l} h(\theta_{li}) p_k(y, \theta_{li})/n_l}{\sum_{i=1}^{n_k} h(\theta_{ki}) p_l(y, \theta_{ki})/n_k}. \tag{9}$$

For the case of independent draws, Meng & Wong (1996) also found the function $h(\theta)$ that minimises the asymptotic relative mean squared error defined, for a generic estimator $\hat{B}_{kl}$ of B_{kl}, by $E(\hat{B}_{kl} - B_{kl})^2/B_{kl}^2$; this optimal choice is

$$h^o(\theta) \propto \{n_k p_k(y, \theta) + n_l p_l(y, \theta) B_{kl}\}^{-1}. \tag{10}$$

Note that computing (10) requires knowledge of B_{kl}, which is obviously unknown. However, given an initial guess for B_{kl}, substituting (10) into (9) allows us to compute an improved estimate of B_{kl}. This procedure, repeated iteratively, converges to the same limit regardless of the initial guess (Meng & Wong, 1996, Theorem 2); asymptotically, the resulting estimator has the same relative mean squared error as that of the optimal one. When the draws from the posteriors are not independent, Meng & Wong (1996) argued that the optimal choice of $h(\theta)$ still has the same form, but with n_k and n_l in (10) replaced with the effective sample sizes.

Note that identity (8) cannot be used when there is no overlap between Θ_k and Θ_l or when the two parameter spaces have different dimensions, and so the approach of Meng

& Wong (1996) has been suitably extended. In particular, for the case of nested models, Chen & Shao (1997b) suggested embedding the lower-dimensional density in the higher-dimensional one by 'patching up' a conditional distribution with known normalising constant so that (8) may be applied directly. The case of densities with non-overlapping supports has been dealt with by Meng & Schilling (2002), who suggest shifting the densities to reduce the distance between them before applying the identity.

Finally note that, as shown by Mira & Nicholls (2004), the estimator of Chib (1995) and Chib & Jeliazkov (2001) belongs to the class of estimators based on the bridge sampling identity (8). This is because $p_k(\bar{\theta}|y)$ may be seen as the ratio of two normalising constants, namely $p_k(y, \bar{\theta})$ of the proposal distribution, when we take $q_k(\theta|\bar{\theta})p_k(y, \bar{\theta})$ as the unnormalised density for this distribution, and $p_k(y)$ of the posterior distribution. Therefore, identity (6), on which the estimator in question is based, may be expressed in terms of (8), with $h(\theta) = \alpha_k(\bar{\theta}, \theta)/p_k(y, \theta)$ used to bridge the two distributions. Note that estimation error comes from the dissimilarity between these distributions. Mira & Nicholls (2004) also suggested using the optimal bridge function $h^o(\theta)$ proposed by Meng & Wong (1996) in order to improve the efficiency of the estimator in question.

2·3. *Estimating posterior model probabilities*

Among the second class of estimation methods of the Bayes factor outlined in § 1, the best known is based on the output of the reversible jump algorithm of Green (1995). To ensure reversibility of the Markov chain on which the reversible jump algorithm is based, we assume that, for any pair of models $(\mathscr{M}_k, \mathscr{M}_l)$, there exists a diffeomorphism $(\theta_l, u_k) = g_{kl}(\theta_k, u_l)$ from $S_{kl} = \{(\theta_k, u_l)\}$ to $S_{lk} = \{(\theta_l, u_k)\}$, where u_k and u_l are suitable vectors of auxiliary variables defined so that S_{kl} and S_{lk} have the same dimension. Therefore, if the current state of the Markov chain is (k, θ_k), a new state, (l, θ_l) say, is proposed by generating u_l from a suitable proposal distribution $q_{l|k}(u_l|\theta_k)$; the proposed move is then accepted with probability

$$\alpha_{kl}(\theta_k, u_l) = \min\{1, \beta_{kl}(\theta_k, u_l)\}, \tag{11a}$$

where

$$\beta_{kl}(\theta_k, u_l) = \frac{p_l(y, \theta_l)q_{k|l}(u_k|\theta_l)}{p_k(y, \theta_k)q_{l|k}(u_l|\theta_k)} J_{kl}(\theta_k, u_l), \tag{11b}$$

in which $J_{kl}(\theta_k, u_l)$ is the Jacobian of the transformation $g_{kl}(\theta_k, u_l)$ and, for simplicity, we have assumed that $p_l = p_k$ and that the probability of proposing $\mathscr{M}_l$ when the current model is $\mathscr{M}_k$ is equal to that of proposing $\mathscr{M}_k$ when the current model is $\mathscr{M}_l$. After a suitable number n, say, of iterations, $p(k|y)$ is estimated as the number of times n_k the chain visited $\mathscr{M}_k$, divided by n; denote this estimator by $\hat{p}(k|y)$. Then the standard estimator of B_{kl} based on the reversible jump output is

$$\hat{B}_{kl}^{\mathrm{RJ}} = \frac{\hat{p}(k|y)}{\hat{p}(l|y)} = \frac{n_k}{n_l}. \tag{12}$$

3. The proposed approach

3·1. *An extension of the bridge sampling identity*

We propose a new class of estimators of the Bayes factor based on the standard reversible jump output. To illustrate this class we first have to introduce a simple extension of the

bridge sampling identity. In practice, we enlarge the parameter space of any model under comparison with the same auxiliary variables defined within the reversible jump framework so that the enlarged spaces have the same dimension and we are back into the 'same-dimensional' setting outlined in § 2·2.

Consider the following conditional distribution of the parameter vector θ_k of $\mathscr{M}_k$ and the vector of auxiliary variables u_l given the observed sample y:

$$p_{kl}(\theta_k, u_l|y) = \frac{f_{kl}(y, \theta_k, u_l)}{p_k(y)}, \quad (\theta_k, u_l) \in S_{kl}, \tag{13}$$

where $f_{kl}(y, \theta_k, u_l) = p_k(y, \theta_k) q_{l|k}(u_l|\theta_k)$ and $q_{l|k}(u_l|\theta_k)$ is the proposal distribution used within the reversible jump algorithm for jumping from $\mathscr{M}_k$ to $\mathscr{M}_l$.

Thus, following the analogy outlined at the end of § 2·2, we can compute B_{kl} as the ratio of the normalising constant of the distribution $p_{kl}(\theta, u|y)$, when we take $f_{kl}(y, \theta, u)$ as the unnormalised density, to that of $p_{lk}(\theta, u|y)$. This intuition leads to the following theorem in which E_{kl} denotes expectation under the first of these distributions and E_{lk} under the second one.

Theorem 1. *For any function $h_{lk}(\theta_l, u_k)$, it holds that*

$$B_{kl} = \frac{p_k(y)}{p_l(y)} = \frac{E_{lk}[f_{kl}\{y, g_{lk}(\theta, u)\} h_{lk}(\theta, u)]}{E_{kl}[f_{lk}\{y, g_{kl}(\theta, u)\} h_{lk}\{g_{kl}(\theta, u)\} J_{kl}(\theta, u)]}, \tag{14}$$

provided the expected values above are bounded away from 0 *and* ∞.

Proof. It is sufficient to consider that the numerator of (14) is equal to

$$\frac{1}{p_l(y)} \int_{S_{lk}} f_{kl}\{y, g_{lk}(\theta, u)\} h_{lk}(\theta, u) f_{lk}(y, \theta, u) d\theta du,$$

which equals, after a change of variables of integration,

$$\frac{1}{p_l(y)} \int_{S_{kl}} f_{kl}(y, \theta, u) h_{lk}\{g_{kl}(\theta, u)\} f_{lk}\{y, g_{kl}(\theta, u)\} J_{kl}(\theta, u) d\theta du.$$

The latter is just $p_k(y)/p_l(y)$ times the denominator of (14). □

On the basis of this result, B_{kl} can be consistently estimated by

$$\frac{\sum_{i=1}^{n_l} f_{kl}\{y, g_{lk}(\theta_{li}, u_{ki})\} h_{lk}(\theta_{li}, u_{ki})/n_l}{\sum_{i=1}^{n_k} f_{lk}\{y, g_{kl}(\theta_{ki}, u_{li})\} h_{lk}\{g_{kl}(\theta_{ki}, u_{li})\} J_{kl}(\theta_{ki}, u_{li})/n_k}, \tag{15}$$

where (θ_{ki}, u_{li}), for $i = 1, \ldots, n_k$, is a sample from $p_{kl}(\theta, u|y)$ and (θ_{li}, u_{ki}), for $i = 1, \ldots, n_l$, is a sample drawn from $p_{lk}(\theta, u|y)$. Note that, when these samples are taken from the reversible jump output, as we suggest, n_k and n_l are not fixed but stochastic since the algorithm randomly jumps between models. Note also that it is not ensured that estimator (15) is better than the standard estimator in (12) based on the reversible jump output for any choice of the bridge function $h_{lk}(\theta, u)$. On the contrary, some simulations not reported here showed that (15) can perform worst than (12) when, for example, we set $h_{lk}(\theta, u)$ equal to the geometric function or the constant function described in Meng & Wong (1996, § 5). Instead, the choices of the bridge function presented in §§ 3·2 and 3·3 lead to a substantial gain of efficiency over the standard reversible jump estimator.

3·2. *Noniterative choice of the bridge function*

A very simple estimator of the Bayes factor can be obtained from (14) by letting

$$h_{lk}(\theta, u) = \frac{\alpha_{lk}(\theta, u)}{f_{kl}\{y, g_{lk}(\theta, u)\}},$$

where $\alpha_{lk}(\theta, u)$ is the acceptance probability of the reversible jump algorithm defined in (11). With this choice, the numerator of (14) simply becomes $E_{lk}\{\alpha_{lk}(\theta, u)\}$, while the denominator becomes

$$E_{kl}\left[\frac{f_{lk}\{y, g_{kl}(\theta, u)\}\alpha_{lk}\{g_{kl}(\theta, u)\}J_{kl}(\theta, u)}{f_{kl}(y, \theta, u)}\right],$$

which is equal to $E_{kl}[\beta_{kl}(\theta, u)\alpha_{lk}\{g_{kl}(\theta, u)\}]$ and, in turn, to $E_{kl}\{\alpha_{kl}(\theta, u)\}$. This is because definition (11) implies that if $\alpha_{lk}\{g_{kl}(\theta, u)\} = 1$ then $\beta_{kl}(\theta, u)$ is less then 1 and it coincides with $\alpha_{kl}(\theta, u)$; if instead $\alpha_{lk}\{g_{kl}(\theta, u)\} < 1$, we have that $\beta_{kl}(\theta, u)\alpha_{lk}\{g_{kl}(\theta, u)\} = 1$, the same value attained by $\alpha_{kl}(\theta, u)$. Therefore, the identity at issue becomes $B_{kl} = E_{lk}\{\alpha_{lk}(\theta, u)\}/E_{kl}\{\alpha_{kl}(\theta, u)\}$ and the Bayes factor may be consistently estimated by

$$\hat{B}^{*}_{kl} = \frac{\sum_{i=1}^{n_l} \alpha_{lk}(\theta_{li}, u_{ki})/n_l}{\sum_{i=1}^{n_k} \alpha_{kl}(\theta_{ki}, u_{li})/n_k}. \tag{16}$$

Note that all the quantities required to compute this estimator are provided by the standard reversible jump output. Therefore, as will be shown through the simulations presented in § 4, this estimator is simply a more efficient way of postprocessing the reversible jump output without extra computational effort. Intuitively, the gain in efficiency is due to the fact that, while $\hat{B}^{\mathrm{RJ}}_{kl}$ is based on an auxiliary random process for jumping from one model to another, which obviously increases the variability of the estimator, our estimator $\hat{B}^{*}_{kl}$ is obtained by integrating out this auxiliary random process and thus it may be seen as a Rao–Blackwellised version of $\hat{B}^{\mathrm{RJ}}_{kl}$; see also Casella & Robert (1996).

3·3. *Iterative choice of the bridge function*

As indicated in § 3·1, our extension of the bridge sampling identity is obtained by enlarging the parameter spaces of the models under comparison so that we can go back to the same-dimensional setting of Meng & Wong (1996). Therefore, from their Theorem 1 it follows that the optimal bridge function, i.e. the one leading to the smallest asymptotic relative mean squared error under independent draws, is

$$h^{o}_{lk}(\theta, u) \propto [n_k f_{kl}\{y, g_{lk}(\theta, u)\} + n_l f_{lk}(y, \theta, u) J_{kl}\{g_{lk}(\theta, u)\} B_{kl}]^{-1}. \tag{17}$$

We denote the resulting estimator of the Bayes factor between $\mathscr{M}_k$ and $\mathscr{M}_l$ by $\hat{B}^{\dagger}_{kl}$.

As stressed in Meng & Wong (1996), the choice of $h_{lk}(\theta, u)$ given in (17) is optimal when the draws from $p_{kl}(\theta, u|y)$ and $p_{lk}(\theta, u|y)$ are independent. In the present case, these draws are obtained from the standard output of the reversible jump algorithm, and therefore they are definitely not independent. Meng & Wong (1996) conjecture that in this situation the optimal bridge function has still the same form as that given under independent draws, but with n_k and n_l being the effective sample sizes. In order to estimate the effective size of the sample from $p_{kl}(\theta, u|y)$, we divided n_k by an estimate of the integrated autocorrelation time, $\tau = \sum_{h=-\infty}^{\infty} \rho_h$, where

$$\rho_h = \mathrm{cov}[f_{kl}\{y, g_{lk}(\theta_{li}, u_{ki})\}, f_{kl}\{y, g_{lk}(\theta_{l,i+h}, u_{k,i+h})\}]/\sigma^2$$

with σ^2 being the variance of $f_{kl}\{y, g_{lk}(\theta, u)\}$ under the stationary distribution. To estimate τ we rely on the adaptive truncated periodogram estimator of Sokal (1989), given by $\hat{\tau} = \sum_{|h| \leqslant M} \hat{\rho}_h$, where the window width M is chosen adaptively as the minimum integer such that $M \geqslant 3\hat{\tau}$. In particular, we use the fast Fourier transform to estimate the autocorrelations ρ_h. With the same procedure we estimate the effective sample size for the sample from $p_{lk}(\theta, u|y)$. We denote by $\hat{B}^{\ddagger}_{kl}$ the estimator of the Bayes factor between models $\mathscr{M}_k$ and $\mathscr{M}_l$ obtained on the basis of (15), using the optimal choice of $h_{lk}(\theta, u)$ given in (17) and adjusting for the effective sample sizes.

Finally note that computation of each of the estimators $\hat{B}^{\dagger}_{kl}$ and $\hat{B}^{\ddagger}_{kl}$ needs an iterative procedure, since the function $h^{o}_{lk}(\theta, u)$ depends on the unknown B_{kl}. Thus $\hat{B}^{\dagger}_{kl}$ and $\hat{B}^{\ddagger}_{kl}$ require some additional, though typically minor, computational time compared to $\hat{B}^{*}_{kl}$.

3·4. *Use of the proposed estimator in practice*

The conventional estimator based on the reversible jump output can be directly applied to estimate B_{kl} for any pair of models $(\mathscr{M}_k, \mathscr{M}_l)$, since it is given by the number of times the Markov chain visited $\mathscr{M}_k$ divided by the number of times it visited $\mathscr{M}_l$. Instead, the estimators we propose can be used directly only when jumps between $\mathscr{M}_k$ and $\mathscr{M}_l$ are allowed. However, the reversible jump algorithm is usually implemented so that, once a given model is reached, it can jump only to a limited number of other models. In the following we show that this does not limit the applicability of our approach.

Suppose that, once $\mathscr{M}_k$ is reached, the reversible jump algorithm can jump only to $\mathscr{M}_{k-1}$, if $k > 1$, or to $\mathscr{M}_{k+1}$, if $k < K$, so that the samples required to compute our estimators $\hat{B}^{*}_{k+1,k}$, $\hat{B}^{\dagger}_{k+1,k}$ and $\hat{B}^{\ddagger}_{k+1,k}$ are available for $k = 1, \ldots, K-1$. Then, for any pair of nonconsecutive models $\mathscr{M}_k$ and $\mathscr{M}_l$, with $l > k$, we can estimate B_{kl} by

$$\hat{B}_{kl} = \hat{B}_{k,k+1}\hat{B}_{k+1,k+2}\ldots\hat{B}_{l-1,l},$$

where $\hat{B}$ indicates any of our estimators. By inverting (3) we may also estimate the posterior probabilities $p(k|y)$; when $p_k = 1/K$ $(k = 1, \ldots, K)$ for simplicity, we have

$$\hat{p}(k|y) = \frac{\hat{B}_{k1}}{1 + \hat{B}_{21} + \hat{B}_{31} + \ldots + \hat{B}_{K1}}, \quad \hat{B}_{k1} = (\hat{B}_{1k})^{-1}.$$

4. SOME APPLICATIONS

4·1. *Linear regression analysis*

Han & Carlin (2001) compared several methods for estimating the Bayes factor between two nonnested linear regression models used to analyse a dataset concerning the maximum compressive strength parallel to the grain, Y, for 42 specimens of radiata pine with density, X, and resin-adjusted density, Z, as possible explanatory variables; see also Carlin & Chib (1995). The two competing models are

$$\mathscr{M}_1: \quad Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, v^2),$$

$$\mathscr{M}_2: \quad Y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \tau^2),$$

with the following prior distributions: $N(3000, 10^6)$ for both α and γ, $N(185, 10^4)$ for both β and δ and $\text{IG}\{3, 1/(2 \times 300^2)\}$ for both v^2 and τ^2, where $\text{IG}(a, b)$ denotes the inverse gamma distribution.

The reversible jump algorithm used in this setting is based on $n = 60\,000$ iterations, of which the first 10 000 are treated as burn-in, and on two types of move, namely within-model and across-models, each with probability $\frac{1}{2}$. As in Han & Carlin (2001), we used the following proposal distributions to update the parameters within $\mathscr{M}_1$:

$$\alpha^* \sim N(\alpha, 5000), \quad \beta^* \sim N(\beta, 250), \quad v^{2*} \sim \text{LN}\{\log(v^2), 1\},$$

where $\text{LN}(\mu, \sigma^2)$ denotes the lognormal distribution. For the parameters γ, δ and τ^2 of $\mathscr{M}_2$ we used, respectively, the same proposals as for α, β and v^2, while, to jump from $\mathscr{M}_1$ to $\mathscr{M}_2$, we simply let $(\gamma^*, \delta^*, \tau^{2*}) = (\alpha, \beta, v^2)$, and similarly to jump from $\mathscr{M}_2$ to $\mathscr{M}_1$. This is the equivalent of taking the functions $g_{12}(\theta_1, u_2)$ and $g_{21}(\theta_2, u_1)$, where $\theta_1 = (\alpha, \beta, v^2)$ and $\theta_2 = (\gamma, \delta, \tau^2)$, as identity functions and auxiliary vectors u_1 and u_2 of null dimension; the Jacobian of the transformation in question is always equal to 1.

To compare in terms of efficiency the estimators of the Bayes factor between $\mathscr{M}_2$ and $\mathscr{M}_1$, whose true value is $B_{21} = 4862$, as given in an unpublished University of Nottingham technical report by P. J. Green and A. O'Hagan, we relied on 100 Monte Carlo simulations. The results of these simulations are displayed in Table 1, where by relative error we mean the square root of the relative mean squared error defined in § 2·2.

Table 1. *Comparison of the Bayes factor estimators for the data in Han & Carlin* (2001) *on the basis of* 100 *Monte Carlo simulations*

	$\hat{B}_{21}^{\text{RJ}}$	$\hat{B}_{21}^{*}$	$\hat{B}_{21}^{\dagger}$	$\hat{B}_{21}^{\ddagger}$
Mean	4671·1	4864·8	4864·3	4848·9
Standard error	1261·7	204·5	204·4	246·3
Relative error	26·25%	4·21%	4·20%	5·07%

According to these results, the use of our estimator $\hat{B}_{21}^{*}$ considerably improves the efficiency in estimating B_{21} relative to $\hat{B}_{21}^{\text{RJ}}$ with no extra computational time: the relative error decreases from 26·25% to 4·21%. The estimator $\hat{B}_{21}^{\dagger}$ has an efficiency very similar to that of $\hat{B}_{21}^{*}$, while $\hat{B}_{21}^{\ddagger}$ performs better than $\hat{B}_{21}^{\text{RJ}}$ but worse than the other two. This depends on the fact that $\mathscr{M}_2$ is much more likely than $\mathscr{M}_1$, and therefore the reversible jump algorithm seldom jumps from $\mathscr{M}_2$ to $\mathscr{M}_1$; in other words, we have a very low acceptance rate for this kind of move and thus a very small sample from $p_{12}(\theta, u|y)$. In this case the estimate of the effective sample size becomes particularly unreliable. Han & Carlin (2001) overcome this problem by letting the priors of the models equal $p_1 = 0{\cdot}9995$ and $p_2 = 0{\cdot}0005$; however, this requires extra programming and computing time.

4·2. *Logistic regression analysis*

Dellaportas et al. (2002) compared several methods for selecting a hierarchical logistic regression model for the number of survivals, Y, in a sample of 79 subjects suffering a certain illness using the patient condition, A, and the received treatment, B, as explanatory factors.

We have five possible models: $\mathscr{M}_1$ (intercept); $\mathscr{M}_2$ (intercept + A); $\mathscr{M}_3$ (intercept + B); $\mathscr{M}_4$ (intercept + A + B); $\mathscr{M}_5$ (intercept + A + B + $A.B$). In particular, the full model, $\mathscr{M}_5$, is formulated as

$$Y_{ij} \sim \text{Bi}(n_{ij}, p_{ij}), \quad \text{logit}(p_{ij}) = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB},$$

where, for $i, j = 1, 2$, Y_{ij}, n_{ij} and p_{ij} are, respectively, the number of survivals, the total number of patients and the probability of survival for the patients with condition i who received treatment j. Dellaportas et al. (2002) also used the sum-to-zero identifiability constraint and the prior $N(0, 8)$ for any of the identifiable parameters, μ, μ_2^A, μ_2^B and μ_{22}^{AB}, which by assumption are also a priori independent. The same assumptions are made for any reduced model. Finally, the following proposal was used to update the parameters within the same model, within-model move, and also to jump from one model to another, across-model move:

$$\mu^* \sim N(-0{\cdot}47, 0{\cdot}27), \quad \mu_2^{A*} \sim N(-0{\cdot}87, 0{\cdot}27),$$
$$\mu_2^{B*} \sim N(0{\cdot}56, 0{\cdot}28), \quad \mu_{22}^{AB*} \sim N(-0{\cdot}17, 0{\cdot}27).$$

Thus, for example, to jump from $\mathscr{M}_2$ to $\mathscr{M}_3$, we use vectors of auxiliary variables, u_2 and u_3, of size 1, with the second one proposed from a $N(0{\cdot}56, 0{\cdot}28)$ distribution. Then $(\theta_3, u_2) = g_{23}(\theta_2, u_3)$ is a function that permutes the elements of (θ_2, u_3) in a suitable way; the Jacobian of this function is again equal to 1.

In this setting we used 30 000 interations, discarding the first 5000 as burn-in, for the reversible jump algorithm. As in § 4·1, the estimators of the Bayes factor have been compared, on the basis of their efficiency, using 100 Monte Carlo simulations; the results of this comparison are shown in Table 2, where as true value of any B_{kl} we took the mean over all simulations and estimators.

Table 2. *Comparison of the Bayes factor estimators for the data in Dellaportas et al.* (2002) *on the basis of* 100 *Monte Carlo simulations*

		B_{21}	B_{32}	B_{43}	B_{54}
$\hat{B}_{kl}^{\mathrm{RJ}}$	Mean	101·3800	0·0233	38·4130	0·1174
	Standard error	10·5180	0·0025	3·8397	0·0033
	Relative error	10·57%	10·97%	9·95%	2·81%
$\hat{B}_{kl}^{*}$	Mean	99·8690	0·0228	39·1240	0·1181
	Standard error	0·6894	0·0003	0·4521	0·0017
	Relative error	0·74%	1·43%	1·29%	1·50%
$\hat{B}_{kl}^{\dagger}$	Mean	99·7830	0·0228	39·0520	0·1176
	Standard error	0·7754	0·0003	0·4954	0·0019
	Relative error	0·85%	1·36%	1·32%	1·60%
$\hat{B}_{kl}^{\ddagger}$	Mean	99·5730	0·0229	39·0470	0·1176
	Standard error	2·7192	0·0006	0·5460	0·0016
	Relative error	2·78%	2·72%	1·45%	1·35%

Again the standard estimator based on the reversible jump algorithm seems to be the least efficient in estimating the Bayes factor. The three estimators that we propose perform much better with very similar relative errors. In particular, note that the estimator $\hat{B}_{kl}^{\ddagger}$ performs better than the other two only in estimating B_{54}, which is close to 1.

5. Discussion

Since computation of $\hat{B}_{kl}^{*}$ is only marginally more complicated than that of the usual estimator $\hat{B}_{kl}^{\mathrm{RJ}}$, and yet the gain in terms of efficiency is consistently high, we recommend its use in practical applications. On the other hand the extra improvement in efficiency

obtained when using the optimal estimator $\hat{B}^{\dagger}_{kl}$ is not always worth the extra computational and programming effort. Likewise, $\hat{B}^{\ddagger}_{kl}$ is not always more efficient than $\hat{B}^{\dagger}_{kl}$.

Theoretically, the fact that $\hat{B}^{*}_{kl}$ performs nearly as well as $\hat{B}^{\dagger}_{kl}$ can be justified by noting that both estimators are based on bridge functions belonging to the same power family defined by Meng & Wong (1996, § 5). In our context, the power family is defined as

$$h_{lk}(\theta, u) = ([f_{kl}\{y, g_{lk}(\theta, u)\}]^{1/a} + [A f_{lk}(y, \theta, u) J_{kl}\{g_{lk}(\theta, u)\}]^{1/a})^{-a},$$

for preselected constants $a > 0$ and $A > 0$. The optimal bridge function is obtained by letting $a = 1$ and $A = B_{kl} n_l / n_k$. Note in particular that, since the samples are drawn from the reversible jump algorithm, n_l/n_k is an estimator of B_{lk} and thus $A \simeq 1$. Instead, the bridge function used in $\hat{B}^{*}_{kl}$ is obtained with $a \to 0$ and $A = 1$, which results in

$$\begin{aligned} h_{lk}(\theta, u) &\to \frac{1}{\max[f_{kl}\{y, g_{lk}(\theta, u)\}, f_{lk}(y, \theta, u) J_{kl}\{g_{lk}(\theta, u)\}]} \\ &= \frac{1}{f_{kl}\{y, g_{lk}(\theta, u)\} \max[1, \beta_{kl}\{g_{lk}(\theta_k, u)\}]} \\ &= \frac{\min\{1, \beta_{lk}(\theta, u)\}}{f_{kl}\{y, g_{lk}(\theta, u)\}} = \frac{\alpha_{lk}(\theta, u)}{f_{kl}\{y, g_{lk}(\theta, u)\}}. \end{aligned}$$

Thus, it can easily be seen that the bridge function for $\hat{B}^{*}_{kl}$ is an approximation of the optimal one and the approximation improves if the Bayes factor between the two models under comparison is far from 1. Finally, the efficiency of $\hat{B}^{\ddagger}_{kl}$ over $\hat{B}^{\dagger}_{kl}$ depends on how precise our estimates of the integrated autocorrelation time are. Typically, if the Bayes factor between the two models under comparison is close to 1, then the reversible jump algorithm jumps between the models more freely, our estimates of the autocorrelation time are fairly reliable and the gain in efficiency of $\hat{B}^{\ddagger}_{kl}$ over $\hat{B}^{\dagger}_{kl}$ may be relevant.

Acknowledgement

We are grateful to the editor and two anonymous referees for their helpful suggestions. We also acknowledge the financial support of grants from Ministero dell'Istruzione, dell'Università e della Ricerca.

References

Carlin, B. P. & Chib, S. (1995). Bayesian model choice via Markov-chain Monte-Carlo methods. *J. R. Statist. Soc.* B **57**, 473–84.

Casella, G. & Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94.

Chen, M. H. & Shao, Q. M. (1997a). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25**, 1563–94.

Chen, M. H. & Shao, Q. M. (1997b). Estimating ratios of normalizing constants for densities with different dimensions. *Statist. Sinica* **7**, 607–30.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Am. Statist. Assoc.* **90**, 1313–21.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Am. Statist. Assoc.* **96**, 270–81.

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statist. Comp.* **12**, 27–36.

Gelman, A. & Meng, X. L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163–85.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty methods. *J. Comp. Graph. Statist.* **10**, 230–48.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

GREEN, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, Ed. P. J. Green, N. L. Hjort and S. Richardson, pp. 179–98. Oxford: Oxford University Press.

HAN, C. & CARLIN, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: a comparative review, *J. Am. Statist. Assoc.* **96**, 1122–32.

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

JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.* **31**, 203–22.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

LAVINE, M. & SCHERVISH, M. J. (1999). Bayes factors: what they are and what they are not. *Am. Statistician* **53**, 119–22.

MENG, X. L. & SCHILLING, S. (2002). Warp bridge sampling. *J. Comp. Graph. Statist.* **11**, 552–86.

MENG, X. L. & WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831–60.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–91.

MIRA, A. & NICHOLLS, G. (2004). Bridge estimation of the probability density at a point. *Statist. Sinica* **14**, 603–12.

SOKAL, A. D. (1989). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne.*

[*Received September* 2004. *Revised September* 2005]