Bridge estimation of the probability density at a point

Antonietta Mira*
Department of Economics
University of Insubria
Via Ravasi 2
21100 Varese, Italy
antonietta.mira@uninsubria.it

Geoff Nicholls
Department of Mathematics
Auckland University
Private Bag 92019
Auckland, New Zealand
nicholls@math.auckland.ac.nz

July 2000, revised September 2003

Bridge estimation, as described by Meng and Wong in 1996, is used to estimate the value taken by a probability density at a point in the state space. When the normalisation of the prior density is known, this value may be used to estimate a Bayes factor. It is shown that the multi-block Metropolis-Hastings estimators of Chib and Jeliazkov (2001) are bridge sampling estimators. This identification leads to estimators for the quantity of interest which may be substantially more efficient.

Keywords: Bayes factor, Marginal likelihood, Markov chain Monte Carlo, Metropolis-Hastings algorithms.

*Corresponding author: Phone: +39 0332 215363, Fax: +39 0332 21 55 09

1 Motivation

In a recent paper, Chib and Jeliazkov (2001) treat the problem of estimating marginal likelihoods. Their interest is motivated by the Bayesian model choice problem, and the estimation of Bayes factors.

Consider Bayesian inference for parameters $\theta \in \mathcal{X}$ given data y, likelihood $p_{Y|\Theta}(y|\theta)$, prior probability density $p_\Theta(\theta)$ and posterior probability density $p_{\Theta|Y}(\theta|y)$. The prior and posterior are given in terms of their unnormalised densities and corresponding normalisations, so that $p_\Theta = c_\Theta f_\Theta(\theta)$ and $p_{\Theta|Y}(\theta|y) = c_{\Theta|Y} f_{\Theta|Y}$, with $f_{\Theta|Y} = p_{Y|\Theta} f_\Theta$, as is usual. Now, following Chib and Jeliazkov (2001), write $m(y) = p_Y(y)$ for the marginal likelihood. Fix some state $\theta^* \in \mathcal{X}$ (Chib and Jeliazkov (2001) give advice on choosing this state). Bayes rule then yields

$$m(y) = \frac{p_{Y|\Theta}(y|\theta^*)p_\Theta(\theta^*)}{p_{\Theta|Y}(\theta^*|y)}. \tag{1}$$

The idea of deriving an estimator for $m(y)$ from this simple relation is attributed, in Raftery (1996), to Julian Besag. If $p_{Y|\Theta}$ and p_Θ can be evaluated directly, Eqn. (1) gives an estimate for the marginal likelihood in terms of an estimate for the posterior probability density at a point. The $p_{\Theta|Y}(\theta^*|y)$-estimator given in Chib and Jeliazkov (2001) is convenient, as it is given in terms of the proposal and acceptance functions of a Metropolis-Hastings Markov chain Monte Carlo algorithm for $p_{\Theta|Y}(\theta^*|y)$ itself.

Chib and Jeliazkov (2001) present estimators for $p_{\Theta|Y}(\theta^*|y)$ derived from the components of single and multiple block MCMC algorithms. In single block MCMC, all the components of θ are updated in a single Metropolis-Hastings step. In multiple-block MCMC, the parameters are divided into blocks, so that $\theta = (\theta_1, \ldots \theta_B)$. Each block parameter θ_i may itself be multivariate. At a MCMC update, a block is chosen, and the parameters in that block are updated together, while parameters in other blocks are kept fixed.

In the next section we step back from $p_{\Theta|Y}(\theta^*|y)$-estimation and the marginal likelihood. We estimate a probability density $p(\theta^*) = cf(\theta^*)$ at a point θ^* in its support. We will see that the single and multiple-block estimators given for $p(\theta^*)$ by Chib and Jeliazkov (2001) belong to the class of bridge sampling estimators considered in Meng and Wong (1996). This result may seem surprising given that bridge sampling is usually presented as a method for estimating normalizing constants. Still, if one considers the fact that the ratio of the normalized to the unnormalised density, at a given point, is the normalizing constant, one realizes that bridge sampling is equally useful for estimating values of a normalized probability density. The bridge in Chib and Jeliazkov (2001) is a sequence of distributions, defined on spaces of increasing dimension, running from a trivial distribution on a single atom, and finishing at the posterior distribution. The bridge sampling estimation theory developed in Meng and Wong (1996) and by subsequent authors, applies to the

Chib and Jeliazkov bridge. Efficiency gains are immediate, as we see in Section 3, where we compare different estimators on a simple Bayesian example. Since the Gibbs sampler is a special case of the Metropolis-Hastings algorithm (as proved in Gelman (1992)), the methods proposed by Chib (1995) are also special cases of bridge sampling (indeed Chib and Jeliazkov 2001 paper was presented as a generalization of Chib 1995 paper). However, unlike Chib and Jeliazkov (2001) which cites the bridge sampling literature but fails to make the right connection, Chib (1995) predates Meng and Wong (1996) and thus its contribution should be acknowledged in that this was one of the first instances in the literature where the issue of estimating normalizing constants and Bayes factors within a Monte Carlo simulation (a Gibbs sampler in particular) was brought to the attention of the scientific community and investigated.

2 Bride estimation of probability density

2.1 Bridge estimation

Meng and Wong (1996) treat the problem of estimating a ratio of normalising constants. They begin with an identity. For $i = 1, 2$, let $p^{(i)}(\theta)$ be probability densities defined on spaces $\mathcal{X}^{(i)}$. Suppose these probability densities are given

in terms of known densities $f^{(i)}$ and corresponding unknown normalising constants $c^{(i)}$, so that $p^{(i)}(\theta) = c^{(i)} f^{(i)}(\theta)$. Let a function $h(\theta)$ be given, satisfying

$$0 < \left| \int_{\mathcal{X}^{(1)} \cap \mathcal{X}^{(2)}} h(\theta) f^{(1)}(\theta) f^{(2)}(\theta) d\theta \right| < \infty. \tag{2}$$

Such a function always exists as long as the intersection of the supports for f^1 and f^2 is non-empty. Let $r = c^{(1)}/c^{(2)}$ and let E_i be an expectation in $p^{(i)}$. Meng and Wong (1996) estimate r using the identity

$$r = \frac{\mathsf{E}_1[f^{(2)}(\theta)h(\theta)]}{\mathsf{E}_2[f^{(1)}(\theta)h(\theta)]}. \tag{3}$$

They choose $h(\theta)$ to minimise mean square error. Suppose that, for $i = 1, 2$, sequences $S^{(i)} = \{\theta^{(i),j}\}_{j=1}^{N^{(i)}}$ of $N^{(i)}$ iid samples $\theta^{(i),j} \sim p^{(i)}$ are available. The samples may be correlated, for example, they may be generated by MCMC. Let $S = \{S^{(1)}, S^{(2)}\}$ and $\hat{r}(S)$ be an estimate of r based on S. The relative mean square error of $\hat{r}(S)$

$$RE^2(\hat{r}) = \frac{\mathsf{E}_S[(\hat{r}(S) - r)^2]}{r^2}$$

depends on h, and on the joint distribution of the samples S on which it is based. Suppose $h = h_O$ minimises $RE^2(\hat{r})$ over all admissible h. Meng and Wong (1996) show that, for iid sampling, $h_O = h_O(f^{(1)}, f^{(2)})$ with

$$h_O(f^{(1)}, f^{(2)}) = \left[r N^{(1)} f^{(1)}(\theta) + N^{(2)} f^{(2)}(\theta) \right]^{-1}. \tag{4}$$

As those authors explain, the presence of r in the proposed estimator is not an obstacle: the iteration defined by

$$\hat{r}_{t+1}(S) = \frac{\frac{1}{N^{(1)}} \sum_{j=1}^{N^{(1)}} \frac{f^{(2)}(\theta^{(1),j})}{\hat{r}_t(S) N^{(1)} f^{(1)}(\theta^{(1),j}) + N^{(2)} f^{(2)}(\theta^{(1),j})}}{\frac{1}{N^{(2)}} \sum_{j=1}^{N^{(2)}} \frac{f^{(1)}(\theta^{(2),j})}{\hat{r}_t(S) N^{(1)} f^{(1)}(\theta^{(2),j}) + N^{(2)} f^{(2)}(\theta^{(2),j})}} \tag{5}$$

converges (usually very rapidly) to an estimator that is asymptotically equivalent to the optimal estimator based on the true r.

When the samples in S_i are not iid, the sample size, $N^{(i)}$, is not defined. Meng and Wong (1996) consider replacing $N^{(i)}$ in Eqn. (5) with the "effective sample size" parameter of the set S_i. There is no one number which gives the effective sample size of MCMC output, since the serial autocorrelations in MCMC samples vary from one output parameter to another. However, as Meng and Wong show, the relative mean square error $RE^2(\hat{r})$ is typically insensitive to the sample size estimate in a wide neighborhood of the optimal value. In the Bayesian-MCMC setting of Section 3, with $f^{(2)}$ the posterior and $f^{(1)}$ the prior, we replace $N^{(2)}$ in Eqn. (5) with $N^{(2)}/\tau_{Y|\Theta}$ where $\tau_{Y|\Theta}$ the integrated autocorrelation time of the likelihood in the sequence $S^{(2)}$, $\tau_{Y|\Theta} = 1 + 2\sum_{m=1}^{\infty} \rho_{Y|\Theta}(m)$, and $\rho_{Y|\Theta}(m)$ is the autocorrelation at lag m of the likelihood function.

2.2 The probability density at a point

We now use Eqn. (3) to estimate a probability density, $p(\theta^*) = cf(\theta^*)$ say, at a point θ^* in its support. Suppose we have a Metropolis Hastings MCMC algorithm, ergodic with unique equilibrium density $p(\theta)$ on $\mathcal{X}$, proposal probability density $q(\theta, \theta')$ (normalised over its second argument, θ') and acceptance probability

$$\alpha(\theta, \theta') = \min\left(1,\ \frac{f(\theta')q(\theta', \theta)}{f(\theta)q(\theta, \theta')}\right).$$

Now, referring to Eqn. (3), choose $f^{(1)} = f^{(1)}_{CJ}$, $f^{(2)} = f^{(2)}_{CJ}$ and $h = h_{CJ}$ where

$$f^{(1)}_{CJ}(\theta) = f(\theta)$$

and

$$f^{(2)}_{CJ} = f(\theta^*)q(\theta^*, \theta),$$

so that $c^{(1)} = c$, $c^{(2)} = 1/f(\theta^*)$ and hence $c^{(1)}/c^{(2)} = p(\theta^*)$. Set

$$h_{CJ} = \alpha(\theta^*, \theta)/f(\theta), \tag{6}$$

and use the detailed balance relation for q, α and f to get

$$p(\theta^*) = \frac{\mathsf{E}_1[\alpha(\theta, \theta^*)q(\theta, \theta^*)]}{\mathsf{E}_2[\alpha(\theta^*, \theta)]}. \tag{7}$$

In Eqn. (7), E_1 is an expectation in $p(\theta)$ and E_2 is an expectation in $q(\theta^*, \theta)$. Let $\hat{p}_{CJ}$ be the estimator derived from Eqn. (7) by replacing the expectations in that equation with Monte Carlo derived empirical estimates.

The estimator $\hat{p}_{CJ}$ is given in Eqn. (9) of Chib and Jeliazkov (2001). It is in the class considered by Meng and Wong (1996). The optimal estimator in that class, for iid Monte Carlo samples, and given $f_{CJ}^{(1)}$ and $f_{CJ}^{(2)}$, takes $h = h_O(f_{CJ}^{(1)}, f_{CJ}^{(2)})$. Let $\hat{p}_{CJMW}$ denote the corresponding iid-optimal $p(\theta^*)$-estimator. We conjecture that, when the Monte Carlo samples are not iid, $\hat{p}_{CJMW}$ is always superior to $\hat{p}_{CJ}$, though we have no proof. In the example we give in Section 3, and in another unpublished example, the estimated variance of $\hat{p}_{CJMW}$ is always smaller than that of $\hat{p}_{CJ}$, as we would expect. For further empirical evidence in this direction see Meng and Schilling (2002).

2.3 Multiple-block estimators

In the multiple-block setting, the multivariate parameter $\theta = (\theta_1, \ldots \theta_B)$ is divided up into B blocks of parameters. A state $\theta^* = (\theta_1^*, \ldots \theta_B^*), \theta^* \in \mathcal{X}$ is chosen. We discuss estimation of probability density $p(\theta^*) = p(\theta_1^*, \ldots \theta_B^*)$ from the component functions of an MCMC algorithm realising $\theta \sim p$. As before, $p(\theta) = cf(\theta)$, with c an unknown normalising constant and f a density function over $\mathcal{X}$.

Chib and Jeliazkov (2001) set up a ladder of densities, fixing one block of parameters at each step. Let $\psi_i^* = (\theta_1^*, \ldots \theta_{i-1}^*)$ and $\psi^i = (\theta_{i+1}, \ldots \theta_B)$ so that, for example, $(\psi_i^*, \theta_i, \psi^i) = (\theta_1^*, \ldots \theta_{i-1}^*, \theta_i, \theta_{i+1}, \ldots \theta_B)$. In multiple-block

MCMC, just one parameter block is updated at each step of the Markov chain. The proposal probability density for block i is $q(\theta_i, \theta_i'|\psi_i, \psi^i)$, and the acceptance probability is $\alpha(\theta_i, \theta_i'|\psi_i, \psi^i)$. Chib and Jeliazkov (2001) start with the identity,

$$p(\theta_1^*, \ldots \theta_B^*) = \prod_{i=1}^{B} p(\theta_i^*|\psi_i^*). \tag{8}$$

They estimate each factor on the right side of Eqn. (8) and take a product.

Each of the factors $p(\theta_i^*|\psi_i^*)$ in (8) can be written as a ratio of normalizing constants as follows. For $i = 1, 2, \ldots B$, let c_i normalize $p(\theta_i, \psi^i|\psi_i^*)$, so that

$$p(\theta_i, \psi^i|\psi_i^*) = c_i f(\psi_i^*, \theta_i, \psi^i).$$

Then

$$p(\theta_i^*|\psi_i^*) = c_i/c_{i+1},$$

a ratio of normalizing constants. The special cases are $i = 1$ where $c_1 = c$ and $i = B$ where

$$p(\theta_B^*|\psi_B^*) = c_B f(\theta_1^*, \ldots \theta_B^*).$$

Relation (8) is just

$$p(\theta_1^*, \ldots \theta_B^*) = \frac{c_1}{c_2} \times \frac{c_2}{c_3} \times \ldots \frac{c_{B-1}}{c_B} \times c_B f(\theta_1^*, \ldots \theta_B^*).$$

Now, referring to (3), the choices

$$f^{(1)} = f(\psi_i^*, \theta_i, \psi^i),$$

$$f^{(2)} = f(\psi_i^*, \theta_i^*, \psi^i) q(\theta_i^*, \theta_i | \psi_i^*, \psi^i),$$

and

$$h = \frac{\alpha(\theta_i^*, \theta_i | \psi_i^*, \psi^i)}{f(\psi_i^*, \theta_i, \psi^i)},$$

lead to

$$\frac{c_i}{c_{i+1}} = \frac{\mathsf{E}_1[\alpha(\theta_i, \theta_i^* | \psi_i^*, \psi^i) q(\theta_i, \theta_i^* | \psi_i^*, \psi^i)]}{\mathsf{E}_2[\alpha(\theta_i^*, \theta_i | \psi_i^*, \psi^i)]}. \tag{9}$$

Eqn. (9) determines a c_i/c_{i+1}-estimator, which Chib and Jeliazkov (2001) use to estimate the $p(\theta_i^*|\psi_i^*)$-factors on the right of Eqn. (8). The resulting multi-block estimator is in the class discussed in Meng and Wong (1996). The iid-optimal estimator for each factor is given by setting $h = h_O(f^{(1)}, f^{(2)})$.

3 Example

3.1 Estimators

In this section we use the $p(\theta^*)$-estimators $\hat{p}_{CJ}$ and $\hat{p}_{CJMW}$ defined in Section 2.3, and single-block MCMC simulation, in order to estimate a Bayesian posterior density $p_{\Theta|Y}(\theta^*|y)$ and marginal likelihood $m(y)$.

Eqn. (1) determines $m(y)$-estimates $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$ using $p_{\Theta|Y}$-estimators $\hat{p}_{CJ}$ and $\hat{p}_{CJMW}$ respectively. In order to compute $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$ we need the prior normalising constant c_Θ. Suppose our calculation of c_Θ is out by a constant multiple. The Bayes factor we form as a ratio of m-

Table 1: The estimators compared in Section 3. Functions $f^{(1)}$, $f^{(2)}$ and h refer to Eqn. (3). In column one, $p^*_{Y|\Theta}p^*_\Theta \equiv p_{Y|\Theta}(y|\theta^*)p_\Theta(\theta^*)$.

Estimator	$f^{(1)}$	$f^{(2)}$	h
$\hat{m}_{MW} = \widehat{c_\Theta/c_{\Theta\|Y}}$	$f_\Theta(\theta)$	$f_{\Theta\|Y}(\theta\|y)$	$h_O(f^{(1)}_{MW}, f^{(2)}_{MW})$
$\hat{m}_{CJ} = p^*_{Y\|\Theta}p^*_\Theta/\hat{p}_{CJ}$	$f_{\Theta\|Y}(\theta\|y)$	$f_{\Theta\|Y}(\theta^*\|y)q(\theta^*,\theta)$	h_{CJ}
$\hat{m}_{CJMW} = p^*_{Y\|\Theta}p^*_\Theta/\hat{p}_{CJMW}$	$f_{\Theta\|Y}(\theta\|y)$	$f_{\Theta\|Y}(\theta^*\|y)q(\theta^*,\theta)$	$h_O(f^{(1)}_{CJ}, f^{(2)}_{CJ})$

estimates will be out by the same factor. We need an m-estimator which is not exposed to these errors, something simple, which does not require model-specific revisions. For that reason we implemented a third m-estimate obtained directly from Eqn. (3) without the use of Eqn. (1). Following, Meng and Wong (1996), set

$$f^{(1)}_{MW}(\theta) = f_\Theta(\theta),$$

$$f^{(2)}_{MW} = f_{Y|\Theta}(y|\theta)f_\Theta(\theta),$$

and $h = h_O(f^{(1)}_{MW}, f^{(2)}_{MW})$ in Eqn. (3). Now $c^{(1)}/c^{(2)} = c_\Theta/c_{\Theta|Y} = m(y)$, so these substitutions determine an m-estimator, $\hat{m}_{MW}$ say. We implement $\hat{m}_{MW}$ using the iteration given in Eqn. (5). The three estimators we consider are laid out in Table 1.

The discussion above illustrates an attractive feature of bridge estimation. We can easily construct several estimators for the same quantity. In partic-

ular, in our setting, there are at least two natural choices: either bridging the Metropolis-Hastings proposal with the target or the posterior with the prior. The latter is usually not efficient since the two are often quite far from each other but has the advantage of avoiding the calculation of the prior normalizing constant. In general, the idea is to design estimators with complementary weaknesses. Thus $\hat{m}_{CJMW}$ is relatively precise but (like $\hat{m}_{CJ}$) potentially inaccurate, whilst m_{MW} is relatively imprecise, but less exposed to inaccuracies in implementation. This feature is visible in our example. See Section 3.2, penultimate paragraph.

3.2 Application

Our example is a posterior density developed by Carlin and Louis (1996) from the flour-beetle mortality data of Bliss (1935). There are three parameters, which may be sampled as a block. In experiment $i = 1, 2, \ldots K = 8$, a_i beetles are exposed to a poison dose of magnitude b_i resulting in c_i deaths. The observation model for the data $y_i = (a_i, b_i, c_i), i = 1, \ldots K$ is a modified logistic model with parameters $\theta = (\mu, \sigma, m)$. The likelihood factor $p_{Y_i|\Theta}(y_i|\theta)$ can be written

$$p_{Y_i|\Theta}(y_i|\theta) = I(b_i, \theta)^{c_i} \left[1 + I(b_i, \theta)\right]^{a_i - c_i},$$

where

$$I(b_i, \theta) = \left[\frac{e^{\frac{b_i - \mu}{\sigma}}}{1 + e^{\frac{b_i - \mu}{\sigma}}} \right]^m .$$

The prior $f_\Theta(\theta)$ is expressed in terms of prior parameters $a_0, b_0, \ldots f_0$,

$$\begin{aligned} \mu &\sim N(a_0, b_0^2), \\ \sigma^2 &\sim \text{I}\Gamma(c_0, d_0), \\ m &\sim \Gamma(e_0, f_0). \end{aligned}$$

The posterior probability density is then

$$p_{\Theta|Y}(\theta|y) = c_{\Theta|Y} f_\Theta(\theta) \prod_{k=1}^{K} p_{Y_i|\Theta}(y_i|\theta).$$

Carlin and Louis (1996) explain why $p_{\Theta|Y}$ may not conveniently be Gibbs-sampled, and give the following Metropolis Hastings algorithm. They make a change of variables $\theta = (\mu, \frac{1}{2}\log(\sigma^2), \log(m))$ and carry out MCMC in the new variables. The proposal density $q(\theta, \theta')$ is multivariate normal about the current state, $\theta' \sim N(\theta, s\hat{\Sigma})$ with $\hat{\Sigma}$ the estimated posterior covariance, and s a free positive parameter.

As we vary s we vary the efficiency of the MCMC. We measure MCMC efficiency using $\tau_{Y|\Theta}$, the integrated autocorrelation time of the likelihood statistic output from MCMC simulation of the posterior. Output of length N updates has an effective sample size $N/\tau_{Y|\Theta}$ for $p_{Y|\Theta}$-estimation. The quantity $\tau_{Y|\Theta}$ is equal to the sum of the autocorrelation function of the likelihood from lag zero up. It is estimated using the monotone sequence estimator of

Geyer (1992). In Figure 1 we give $\hat{\tau}_{Y|\Theta}$ as a function of s. The MCMC is most efficient for s values around one.

3.3 Results

Since $\hat{m}_{CJMW}$ is certainly optimal given $f_{CJ}^{(1)}$, $f_{CJ}^{(2)}$ and iid simulation, we should include *inefficient* MCMC samplers in our experiment. As we vary MCMC parameter s in the range $0.05 \leq s \leq 20$, effective sample size for $p_{Y|\Theta}$-estimation varies by over an order of magnitude. We estimate the variances of $\hat{m}_{CJ}, \hat{m}_{CJMW}$ and $\hat{m}_{MW}$ using multiple independent estimates at each s value. For each s value we made $K = 30$ independent MCMC simulations of length $N = 10000$ updates from $p_{\Theta|Y}$ (10 s-values so 300 runs each of length 10^4). The effective sample size for the likelihood statistic in a single run varied from about 500 down to about 20. For each of the K simulations associated with each s-value we compute a $\hat{m}_{MW}$, a $\hat{m}_{CJ}$ and a $\hat{m}_{CJMW}$. The θ^* value used in computing $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$ from a given MCMC run is chosen to be the state of highest posterior density encountered in that run. N samples from $q(\theta^*, \theta)$ are drawn independently for each run, in order to compute $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$. Similarly N samples from p_Θ are drawn independently for each run, in order to compute $\hat{m}_{MW}$. As discussed at the ends of Sections 2.1 and 3.2, we choose to evaluate Eqn. (5) using $N^{(2)} = N/\hat{\tau}_{Y|\Theta}$. The $\hat{\tau}_{Y|\Theta}$-value used for each $\hat{m}_{MW}$ and $\hat{m}_{CJMW}$ estimate is

formed from the same output used to compute $\hat{m}_{MW}$ and $\hat{m}_{CJMW}$ themselves.

The main results are summarized in Figure 2. The salient features of Figure 2 are, first, the estimator $\hat{m}_{CJMW}$ achieves a lower variance than estimator $\hat{m}_{CJ}$. The gain is above an order of magnitude where the MCMC is tuned so that its updates are efficient. Secondly, $\hat{m}_{CJMW}$ and $\hat{m}_{CJ}$ require the knowledge of the prior and likelihood normalizing constants. The estimator $\hat{m}_{MW}$ does not use that information. Unsurprisingly it is less efficient. On the other hand one is not required to make accurate hand-calculations of prior normalizations where $\hat{m}_{MW}$ is used. For that reason it is good practice to compute $\hat{m}_{MW}$ when $m(y)$ is needed in practical applications as a check on $\hat{m}_{CJMW}$.

Our MCMC implementation was checked by making comparisons with data (for example, $\hat{\Sigma}$) published in Carlin and Louis (1996). Marginal likelihood estimates at $s = 0.37$, combining results from 30 runs, were $\hat{m}_{CJ} = (1.518 \pm 0.014) \times 10^{-84}$, $\hat{m}_{CJMW} = (1.521 \pm 0.004) \times 10^{-84}$ and $\hat{m}_{MW} = (1.9 \pm 0.4) \times 10^{-84}$ at 95% confidence.

4 Conclusions

We have shown that the multi-block Metropolis-Hastings estimators of Chib and Jeliazkov (2001) are bridge sampling estimators. By choosing the free

functions in those general bridge estimation identities appropriately we arrive at the identity Chib and Jeliazkov (2001) use to define their estimators. Results in Meng and Wong (1996) and Chen and Shao (1997) lead to efficiency gains in estimation. Simulation results on very simple single block problems confirm these gains. We emphasise that the estimates $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$ reported in Section 3.3 are computed from exactly the same sample sets.

We may use different MCMC proposal densities, q^{SIM} and q^{EST} say, in simulation and estimation. This point is made in Chib and Jeliazkov (2001) and can be addressed formally *via* the Hellinger distance. Meng and Wong (1996) show that, for iid-samples, the variance of $\widehat{c^{(1)}/c^{(2)}}$ is sensitive to the Hellinger distance between $f^{(1)}$ and $f^{(2)}$, and hence to q^{EST}. What happens if we switch to MCMC sampling, keep q^{EST} fixed and vary q^{SIM}? Does the same distance control the variance of $\widehat{c^{(1)}/c^{(2)}}$? These remain open questions.

Chen and Shao (1997) extend Meng and Wong (1996) to situations in which $f^{(1)}$ and $f^{(2)}$ do not overlap. They use a transition kernel to link the two spaces. The estimators in Chib and Jeliazkov (2001) can be expressed in that setting, with q playing the role of the link. In our presentation the link function $q(\theta_i^*, \theta_i)$ is absorbed into $f^{(2)}$, adding the dimensions needed to put $f(\psi_i^*, \theta_i^*, \psi^i)q(\theta_i^*, \theta_i|\psi_i^*, \psi^i)$ on the same space as $f(\psi_i^*, \theta_i, \psi^i)$. The whole framework can be expressed in terms of distributions and the reversible jump

Monte Carlo of Green (1995). It should be noted that the usefulness of Chen and Shao's (1996) strategy depends on applications. In some cases it would be better not to try to directly match densities of different dimentions, but rather match each of them to a convenient approximation on the same space, estimate the normalizing constants separately using bridge sampling and then take the ratio. See Section 1 of Meng and Schilling (2002) for a discussion of this issue.

The comparison made in Section 3 was repeated on a radiocarbon model comparison of the kind discussed in Nicholls and Jones (2001). Qualitatively similar results were obtained. $\hat{m}_{CJMW}$ had lower variance than $\hat{m}_{CJ}$ in all measurements. Where there is high serial correlation in MCMC sample output, there is little to separate $\hat{m}_{CJ}$ and $\hat{m}_{CJMW}$. As the MCMC was tuned to become more efficient (or sample spacing increased) the margin in favor of the iid-optimal $\hat{m}_{CJMW}$ grew. For multi-block problems these gains are likely to be more strongly marked, since those bridge sampling estimates involve products. Small efficiency gains accumulate from one factor to the next.

References

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology 22*, 134–167.

Carlin, B. and T. Louis (1996). *Bayes and Empirical Bayes methods for Data analysis*. London: Chapman and Hall.

Chen, M.-H. and Q.-M. Shao (1997). Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica 7*, 607–630.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

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

Gelman, A. (1992). Iterative and non-iterative simulation algorithms. *Computing Science and Statistics 24*, 433–438.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci. 7*, 473–511.

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

Meng, X.-L. and S. Schilling (2002). Warp bridge sampling. *Journal of Computational & Graphical Statistics 11*, 552–586.

Meng, X.-L. and W. Wong (1996). Simulating ratios of normalising constants via a simple identity: a theoretical exploration. *Statistica*

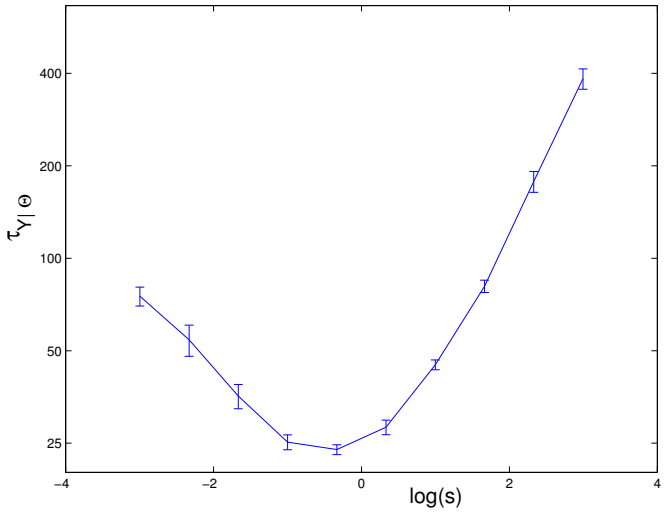

Figure 1: Estimated autocorrelation time $\tau_{Y|\Theta}$ for the likelihood in MCMC simulation of the posterior plotted with MCMC parameter s varying. Note the log-scale at left.

Sinica 6, 831–860.

Nicholls, G. K. and M. Jones (2001). Radiocarbon dating with temporal order constraints. *J. R. Statist. Soc. C 50*, 503–521.

Raftery, A. (1996). *Markov Chain Monte Carlo in Practice*, Chapter 10, pp. 163–187. Chapman and Hall.

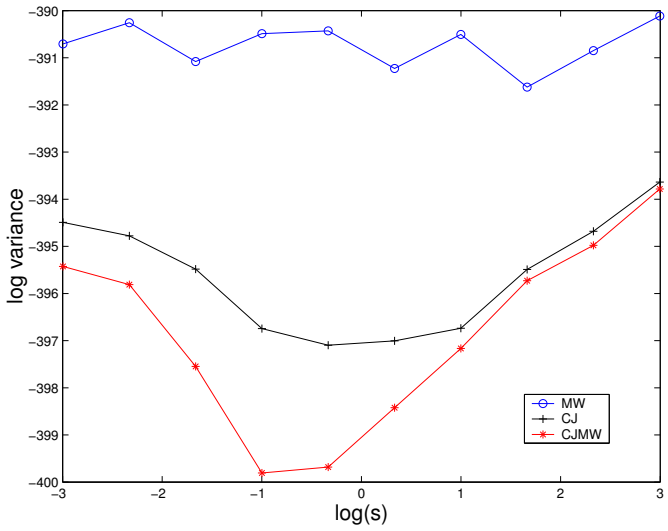

Figure 2: Estimated variance of three different estimators for $m(y)$ as a function of a MCMC-parameter s.