View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

Francesco Bartolucci, Antonietta Mira

Efficient estimate of Bayes factors from Reversible Jump output

2003/33

UNIVERSITÀ DELL'INSUBRIA
FACOLTÀ DI ECONOMIA

http://eco.uninsubria.it

In questi quaderni vengono pubblicati i lavori dei docenti della Facoltà di Economia dell’Università dell’Insubria. La pubblicazione di contributi di altri studiosi, che abbiano un rapporto didattico o scientifico stabile con la Facoltà, può essere proposta da un professore della Facoltà, dopo che il contributo sia stato discusso pubblicamente. Il nome del proponente è riportato in nota all'articolo. I punti di vista espressi nei quaderni della Facoltà di Economia riflettono unicamente le opinioni degli autori, e non rispecchiano necessariamente quelli della Facoltà di Economia dell'Università dell'Insubria.

These Working papers collect the work of the Faculty of Economics of the University of Insubria. The publication of work by other Authors can be proposed by a member of the Faculty, provided that the paper has been presented in public. The name of the proposer is reported in a footnote. The views expressed in the Working papers reflect the opinions of the Authors only, and not necessarily the ones of the Economics Faculty of the University of Insubria.

© Copyright Francesco Bartolucci, Antonietta Mira
Printed in Italy in October 2003
Università degli Studi dell'Insubria
Via Ravasi 2, 21100 Varese, Italy

All rights reserved. No part of this paper may be reproduced in any form without permission of the Author.

Efficient estimate of Bayes factors from Reversible Jump output

Francesco Bartolucci* and Antonietta Mira †

October 27, 2003

Abstract

We extend Meng and Wong (1996) identity from a fixed to a varying dimentional setting. The identity is a very powerful tool to estimate ratios of normalizing constants and thus can be used to evaluate Bayes factors. The extention is driven by the reversible jump algorithm so that the output from the sampler can be directly used to efficiently estimate the required Bayes factor. Two applications, involving linear and logistic regression models, illustrate the advantages of the suggested approach with respect to alternatives previously proposed in the literature.

Keywords: Bayes factor; Bayesian model choice; Marginal likelihood; Markov chain Monte Carlo; Reversible jump.

*Istituto di Scienze Economiche, Unversità di Urbino, Via Saffi, 42, 61029 Urbino, Italy, *email*: Francesco.Bartolucci@uniurb.it

†Dipartimento di Economia, Università dell'Insubria, Via Ravasi, 2, 21100 Varese, Italy, *email*: antonietta.mira@uninsubria.it

1 Introduction

The Bayes factor (BF), defined as the ratio of the marginal likelihoods for a pair of models (see Jeffreys, 1935 and 1961, Kass and Raftery, 1995 and Lavine and Schervish, 1999), represents the evidence provided by the data in favor of a certain model. This is the natural and most widespread model choice criterion in a Bayesian context. Unfortunately, direct computation of the BF is almost always infeasible and so its estimation has attracted considerable interest in the recent Markov chain Monte Carlo (MCMC) literature; good reviews are Dellaportas *et al.* (2001), Han and Carlin (2001) and Green (2003).

Among the numerical methods to estimate the BF proposed in the literature, two have had great success: the Reversible jump (RJ) of Green (1995) and the method of Chib (1995), further extended by Chib and Jelaizkov (2001), which will be indicated by M. The first method requires the Markov chain to be defined over the model and parameter space jointly. This approach delivers the posterior probabilities of each model, as frequency of visits, and an empirical posterior distribution of the model parameters. The approach of Chib (1995), instead, aims at estimating the BF from the output of separate MCMC simulations conducted within each model. Both methods present certain drawbacks. The RJ method requires careful tuning of the proposal distributions to jump between spaces of different dimensions and is very inefficient when one model is decisively better than the others. On the other hand, the M method requires a fair amount of "bookkeeping" and becomes impractical if the number of candidate models is very large.

In the literature on the estimation of the BF, a fundamental contribute is represented by the identity of Meng and Wong (1996) on the basis of which it is possible to estimate the ratio between the normalising constants of two distributions; the resulting estimator is referred to as Bridge estimator. The natural use of this identity is when the two distributions are defined on state spaces with the same dimension. The extension to the case of spaces with different dimension has been attempted by Chen and Shao (1997b) and Mira and Nicholls (2003). The latter, in particular, proves that the estimator of Chib (1996) is a particular instance of the Bridge estimator and shows how its efficiency may be increased through the optimal criterion of Meng and Wong (1996); the resulting estimator will be denoted by M-MW. Chen and Shao (1997b), instead, show how the identity at issue may be used when the state space of one distribution is a subset of that of the other one (nested models). However, their approach requires, ideally (to maximize efficiency), to know the conditional distribution of one subset of the parameters given the rest and thus it is often difficult to implement efficiently.

In this paper we introduce an approach based on the Meng and Wong (1996) identity to estimate the ratio between the normalising constants of two distributions defined on state spaces with different dimension (not necessarily nested). Our approach artificially enlarges, via auxiliary variables, the state space of both distributions so that, at the end, the distributions of the original and the auxiliary variables live on the same dimentional state space. To do this we adopt the strategy suggested by Green (1995) to implement the RJ algorithm that is, the indirect specification of the MCMC proposal distributions as deterministic functions of auxiliary underlying random variables. The same random variables are used, in our setting, to enlarge the state spaces. So, in a way, we generalize the original identity introduced by Meng and Wong (1996) just like the RJ acceptance probability generalizes the original Metropolis-Hastings one.

The proposed approach is more general than that of Chen and Shao (1997b); it is also simpler to use since we can exploit methods to find efficient proposal distributions already developed in the literature on RJ (see Brooks, Giudici and Roberts, 2003, and the references therein). On the other hand, the proposed idea, hereafter denoted by RJ-MW, represents an optimised version of the one recently proposed by Bartolucci and Scaccia (2003) who did not realize that their estimator could be seen as a particular case of the Bridge estimator.

The paper is organized as follows. In Section 2 we introduce some preliminary notation and briefly describe the approaches of Green (1995), Chib and Jelaizkov (2001) and the identity of Meng and Wong (1996). Our approach and its implementation are illustrated in Section 3, while, in Section 4, two applications are presented, involving linear and logistic models. The examples show the advantages of the RJ-MW estimator: at least in the setting studied, the proposed estimator of the BF is more efficient than the alternatives considered.

2 Preliminaries

Let $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$ be a collection of K models, with model $\mathcal{M}_k$ characterized by the parameter vector $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k$ of dimension d_k and let $p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)$ be the likelihood of model $\mathcal{M}_k$. In a Bayesian prospective, for each model $\mathcal{M}_k$, we assign a prior distribution on $\boldsymbol{\theta}_k$, $p(\boldsymbol{\theta}_k|k)$, and denote by $p(k)$ the prior probability of the model. The BF between two models, say $\mathcal{M}_l$ and $\mathcal{M}_k$, is

$$B_{lk} = \frac{p(\boldsymbol{y}|l)}{p(\boldsymbol{y}|k)}, \tag{1}$$

where

$$p(\boldsymbol{y}|k) = \int_{\boldsymbol{\Theta}_k} p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)p(\boldsymbol{\theta}_k|k)d\boldsymbol{\theta}_k$$

is the marginal likelihood for model $\mathcal{M}_k$. The larger is B_{lk}, the greater is the evidence provided by the data in favor of $\mathcal{M}_l$ with respect to $\mathcal{M}_k$. An alternative expression for the BF is

$$B_{lk} = \frac{p(l|\boldsymbol{y})}{p(k|\boldsymbol{y})} \Big/ \frac{p(l)}{p(k)}, \tag{2}$$

where

$$p(k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k)p(k)}{\sum_{h=1}^{K} p(\boldsymbol{y}|h)p(h)} \tag{3}$$

is the posterior probability of model $\mathcal{M}_k$. In almost all real applications exact evaluation of the marginal likelihood is impossible. The RJ method tries to estimate $p(k|\boldsymbol{y})$ for any k and then to estimate the BF between pairs of models on the basis of (2). The M method, instead, aims at directly estimating the marginal likelihood $p(\boldsymbol{y}|k)$ and the BF on the basis of (1).

2.1 Reversible Jump MCMC

The RJ algorithm (Green, 1995) is a natural evolution of the well known Metropolis-Hastings algorithm, which generates observations from the posterior distribution $p(\boldsymbol{\theta}_k, k|\boldsymbol{y})$. To ensure reversibility of the Markov chain running on the joint model and parameter space, for any pair of models $(\mathcal{M}_k, \mathcal{M}_l)$, a bijection is defined, $(\boldsymbol{\theta}_l, \boldsymbol{u}_k) = g_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l)$, from $S_{kl} = \{(\boldsymbol{\theta}_k, \boldsymbol{u}_l)\}$ to $S_{lk} = \{(\boldsymbol{\theta}_l, \boldsymbol{u}_k)\}$, where S_{kl} and S_{lk} have the same dimension. So, if the current state of the Markov chain is $(k, \boldsymbol{\theta}_k)$, a new state, say $(l, \boldsymbol{\theta}_l)$, is proposed by generating auxiliary variables $\boldsymbol{u}_l$ from a suitable *proposal distribution* $q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)$. The proposed move is then accepted with probability

$$\alpha_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l) = \min\left\{1, \frac{p(\boldsymbol{y}, \boldsymbol{\theta}_l|l)p(l)q(k|l)q(\boldsymbol{u}_k|\boldsymbol{\theta}_l)}{p(\boldsymbol{y}, \boldsymbol{\theta}_k|k)p(k)q(l|k)q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)} J_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l)\right\}, \tag{4}$$

where $q(l|k)$ is the probability of proposing model $\mathcal{M}_l$ when the current model is $\mathcal{M}_k$ and J_{kl} is the jacobian of the bijection from S_{kl} to S_{lk}. After a suitable number of iterations, say N, $p(k|\boldsymbol{y})$ is estimated as

$$\hat{p}(k|\boldsymbol{y}) = \frac{n_k}{N},$$

where n_k is the number of times the Markov chain visited model $\mathcal{M}_k$. The algorithm also includes "fixed-dimentional" moves, whereby the model remains the same, although values of the model parameters may be changed.

The RJ algorithm offers a single logical and computational framework for joint inference about the models and the parameters. On the other side, some authors have deemed RJ methods cumbersome to construct and difficult to tune. Moreover, as shown by Bartolucci and Scaccia (2003) the expected value of α_{kl} under the distribution $f(\boldsymbol{\theta}_k, \boldsymbol{u}_l) = p(\boldsymbol{\theta}_k|\boldsymbol{y}, k)q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)$ is limited above by $B_{lk}p(l)/p(k)$. As a consequence, when a model is much more likely than the others, the algorithm gets stuck on that particular model, never visiting the others; so, we are not able to estimate the BF with any degree of accuracy through the RJ method. In principle, as a remedial measure, we could suitably change the priors of the models, but this would require a preliminary estimation of the BFs (for further comments on this technique see Richardson and Green, 1997 and Han and Carlin, 2001, sec. 4.2).

2.2 Meng-Wong identity

Meng and Wong (1996) consider a class of estimators of the ratio of the normalizing constants of two distributions and give the estimator which, under certain conditions, is the most efficient in the class. 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 function $f^{(i)}$ and corresponding unknown normalising constants $c^{(i)}$, so that $p^{(i)}(\theta) = c^{(i)} f^{(i)}(\theta)$. Assume that the two densities have overlapping supports, $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$. Let a function $h(\theta)$ be given, satisfying

$$0 < \left| \int_{\mathcal{X}^{(1)} \cap \mathcal{X}^{(2)}} h(\theta) f^{(1)}(\theta) f^{(2)}(\theta) d\theta \right| < \infty \tag{5}$$

assuming such a function exists. 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

$$r = \frac{\mathsf{E}_1[f^{(2)}(\theta)h(\theta)]}{\mathsf{E}_2[f^{(1)}(\theta)h(\theta)]}. \tag{6}$$

The optimal $h(\theta)$ is chosen to minimise the 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. 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

$$h_O(f^{(1)}, f^{(2)}) = \left[rN^{(1)} f^{(1)}(\theta) + N^{(2)} f^{(2)}(\theta) \right]^{-1}. \tag{7}$$

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

$$\hat{r}_{t+1}(S) = \frac{\frac{1}{N^{(1)}} \sum_{j=1}^{N^{(1)}} \frac{f^{(2)}(\theta^{(1),j})}{\hat{r}_t(S) N^{(1)} f^{(1)}(\theta^{(1),j}) + N^{(2)} f^{(2)}(\theta^{(1),j})}}{\frac{1}{N^{(2)}} \sum_{j=1}^{N^{(2)}} \frac{f^{(1)}(\theta^{(2),j})}{\hat{r}_t(S) N^{(1)} f^{(1)}(\theta^{(2),j}) + N^{(2)} f^{(2)}(\theta^{(2),j})}} \tag{8}$$

converges to $\hat{r}(S)$ (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. (8) 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 setting typically $f^{(2)}$ is the posterior, $f^{(1)}$ the prior. Let $\tau_{Y|\Theta}$ be the integrated autocorrelation time of the likelihood in the sequence $S^{(2)}$, we then replace $N^{(2)}$ in Eqn. (8) with $N^{(2)}/\tau_{Y|\Theta}$.

The case of densities with not overlapping supports is treated in Meng and Schilling (2002) by suggesting to shift the densities to reduce the distance/difference between them before applying identity (6). General class of transformations including centering and stochastic transformations are proposed.

Chen and Shao (1997b) explicitly treat the case of estimating BF when the densities have different dimentions. They only consider the case of nested models and suggest to embed the lower dimensional density into the higher one by "patching up" a conditional distribution with known normalizing constant. Then the identity in Meng and Wong (1996) can be directly applied. Chen and Shao (1997b) give the optimal "patch up" distribution which is the conditional distribution implied by the one with larger dimension.

As commented in Meng and Schilling (2002) weather this approach is better than matching each one of the densities to an approximation on the same space, depends on the application at hand.

3 The proposed extension

Contrary to what suggested in Chen and Shao (1997b) we propose to enlarge the original state spaces with auxiliary variables and embed both densities into a larger dimentional space. The construction of the enlarged state space and the auxiliary variables used are directly derived from the RJ algorithm implemented to sample from the distributions of interest.

Following Bartolucci and Scaccia (2003), consider the distribution of $(\boldsymbol{\theta}_k, \boldsymbol{u}_l)$, with support S_{kl}, already introduced in Section 2.1,

$$p_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l) = f_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l)/c_k, \quad \text{where} \quad f_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l) = p(\boldsymbol{y}, \boldsymbol{\theta}_k|k)q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)$$

Let E_{kl} denote the expected value with respect to p_{kl}, similarly for E_{lk}.

Theorem 1 *For any function $h(\theta_k, u_l)$ with support S_{kl} the following identity holds*

$$\frac{E_{kl}\left[f_{lk}\{g_{kl}(\theta_k, u_l)\}h(\theta_k, u_l)\right]}{E_{lk}\left[f_{kl}\{g_{lk}(\theta_l, u_k)\}h\{g_{lk}(\theta_l, u_k)\}J_{lk}(\theta_l, u_k)\right]} = \frac{c_l}{c_k} = B_{kl}, \tag{9}$$

provided the denominator is bounded away from zero and infinity.

Proof It is sufficient to consider that the numerator of (9) is equal to

$$c_l \int_{S_{kl}} f_{lk}\{(g_{kl}(\theta_k, u_l))\}h(\theta_k, u_l)f_{kl}\{(\theta_k, u_l)\}d\theta_k du_l$$

which equals, after a change of integration variables,

$$c_l \int_{S_{lk}} f_{lk}\{(\theta_l, u_k)\}h\{g_{lk}(\theta_l, u_k)\}f_{kl}\{g_{lk}(\theta_l, u_k)\}J_{lk}(\theta_l, u_k)d\theta_l du_k$$

The latter is nothing but c_l/c_k times the denominator of (9). □

Notice that different functions h might be used for different pairs of models.

Identity 9 implies that the BF can be consistently estimated by

$$\hat{B}_{kl} = \frac{\sum_{i=1}^{N_1} f_{lk}\{g_{kl}(\theta_k^{(i)}, u_l^{(i)})\}h(\theta_k^{(i)}, u_l^{(i)})/N_1}{\sum_{i=1}^{N_2} f_{kl}\{g_{lk}(\theta_l^{(i)}, u_k^{(i)})\}h(\theta_l^{(i)}, u_k^{(i)})J_{lk}(\theta_l^{(i)}, u_k^{(i)})/N_2}, \tag{10}$$

where $(\boldsymbol{\theta}_k^{(i)}, \boldsymbol{u}_l^{(i)})$, $i = 1, \ldots, N_1$, is a sample of dimension N_1 from $f_{kl}(\boldsymbol{\theta}_k, \boldsymbol{u}_l)$ and $(\boldsymbol{\theta}_l^{(i)}, \boldsymbol{u}_k^{(i)})$, $i = 1, \ldots, N_2$, is a sample of dimension N_2 drawn from $f_{lk}(\boldsymbol{\theta}_l, \boldsymbol{u}_k)$. In practice the first sample may be drawn by generating $\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(N_1)}$ from the posterior distribution $p(\boldsymbol{\theta}_k|\boldsymbol{y}, k)$, possibly through an MCMC algorithm, and, for any $\boldsymbol{\theta}_k^{(i)}$, generating $\boldsymbol{u}_l^{(i)}$ from $q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)$; the second sample may be drawn in a similar way. However, for computational efficiency, Bartolucci and Scaccia (2003), propose the following computational algorithm that can be also used in our setting. For any model $\mathcal{M}_k$ generate $\boldsymbol{\theta}_k$ from the posterior distribution $p(\boldsymbol{\theta}_k|\boldsymbol{y}, k)$; then set $k = 1$ and $d = 1$ and perform, a suitable number of times (N), the following operations:

1. if $k > 1$, generate $\boldsymbol{u}_l$ from the proposal $q(\boldsymbol{u}_l|\boldsymbol{\theta}_k)$ and compute $h(\boldsymbol{\theta}_k, \boldsymbol{u}_l)$, for $l = k - 1$;

2. provided that $k < K$, repeat the previous operations for $l = k + 1$;

3. set $k = k + 1$ or $k = k - 1$, according to whether $d = 1$ or $d = 0$ and draw a new value of $\boldsymbol{\theta}_k$ from $p(\boldsymbol{\theta}_k|\boldsymbol{y}, k)$;

4. if $k = K$ set $d = 0$; instead, if $k = 1$, set $d = 1$.

Note that, when $1 < k < K$, we have to compute both $\alpha_{k,k-1}(\boldsymbol{\theta}_k, \boldsymbol{u}_{k-1})$ and $\alpha_{k,k+1}(\boldsymbol{\theta}_k, \boldsymbol{u}_{k+1})$. However, according to (4), these have in common $p(\boldsymbol{y}, \boldsymbol{\theta}_k|k)$ that, consequently, has to be computed only once saving simulation time; this motivates the particular structure of the algorithm. As output of the algorithm we obtain, for any pair of consecutive models, $(\mathcal{M}_k, \mathcal{M}_{k+1})$, the samples required to estimate $B_{k+1,k}$ according to (10). Consequently, we can estimate the BF between any two models, say $\mathcal{M}_l$ and $\mathcal{M}_k$, with $l > k$, as

$$\hat{B}_{lk} = \hat{B}_{l,l-1}\hat{B}_{l-1,l-2}\cdots\hat{B}_{k+1,k}. \tag{11}$$

By inverting (2) we may also estimate the posterior probabilities $p(k|\boldsymbol{y})$'s; when $p(k) = 1/K$, $\forall k$, for simplicity, we have

$$\hat{p}(k|\boldsymbol{y}) = \frac{\hat{B}_{k1}}{1 + \hat{B}_{21} + \hat{B}_{31} + \cdots + \hat{B}_{K1}}. \tag{12}$$

As output of the algorithm, we also obtain, for any model $\mathcal{M}_k$, a sample from the posterior distribution of $\boldsymbol{\theta}_k$. This is a feature in common with the RJ algorithm, the main difference being that, within the algorithm suggested by Bartolucci and Scaccia (2003), these samples have the same dimension, equal to N/K, for all models.

The proposed identity, together with the above mentioned computational algorithm, should allow us to estimate the BF between any pair of models more precisely than the RJ algorithm, especially when one model is much more likely than the others. Intuitively, this is due to the fact that the latter is based on an auxiliary random process for jumping from a model to another, which increases the variability of the estimate. The apparent drawback of computational algorithm is that it may be exploited only when we have a limited number of competing models, whereas, in principle, the RJ algorithm may be used also in presence of a huge number of models.

The main advantange of the computational algorithm suggested by Bartolucci and Scaccia (2003) with respect to the M algorithm of Chib and Jelaizkov (2001) is that it makes use of $2(K-1)$ samples of across-models acceptance probabilities, while the latter is based on $2K$ samples of within-model acceptance probabilities. Moreover, because of its particular structure, the algorithm we adopt is faster in producing these samples.

It may seem that the proposed approach is considerably more difficult to implement than that of Chib and Jelaizkov (2001) and Mira and Nicholls (2003), as these makes use of within-model proposals which, usually, are readily available. However, in many situations, also across-model proposals are readily available (see for instance sec. 4.1), or may be easily obtained from within-model proposals, as when these do not depend on the previous value of the parameter vector (see for instance sec. 4.2). In any case, consider that also a within-model proposal that depends on the previous value of the parameter vector, say $\boldsymbol{\theta}_k$, may be made independent of this by substituting $\boldsymbol{\theta}_k$ with an appropriate $\bar{\boldsymbol{\theta}}_k$ chosen as within the approach of Chib and Jelaizkov (2001).

3.1 Asymptotically optimal choice of h and iterative formula

Once the densities of the two models under comparison are embedded into a common larger state space we are back into the "same dimentional" setting of the paper by Meng and Wong (1996) and the original identity can be applied by identifying appropriate $f^{(1)}$ and $f^{(2)}$. In particular, by taking $\mathcal{X}^{(1)} = \mathcal{X}^{(2)} = (\theta_k, u_l)$ and

$$f^{(1)} = f_{kl}\{g_{lk}(\theta_l, u_k)\}J_{lk}, \quad f^{(2)} = f_{lk}\{g_{kl}(\theta_k, u_l)\},$$

Meng-Wong identity, (6), gives the same result as in (9). Having noticed this, we can used MW asymptotically optimal choice of h:

$$h_0 = \frac{1}{s_l p_{lk} + s_k p_{kl} J_{lk}} \tag{13}$$

(which depends on the unknown ratio) and also adapt to our setting the iterative formula (8).

4 Some applications

In the sequel we compare the proposed method (RJ-MW) to estimate BF with competing ones: the plain RJ, Chib method (M) and Chib method improved using the optimality result by Meng and Wong (1996) (M-MW).

4.1 Linear regression analysis

Han and Carlin (2001) compared several methods for estimating the BF between two non-nested linear regression models used to analyse the data shown in the Table 1, taken from Williams (1959); see also Carlin and Chib (1995).

i	y_i	x_i	z_i	i	y_i	x_i	z_i	i	y_i	x_i	z_i
1	3040	29,2	25,4	15	2250	27,5	23,8	29	1670	22,1	21,3
2	2470	24,7	22,2	16	2650	25,6	25,3	30	3310	29,2	28,5
3	3610	32,3	32,2	17	4970	34,5	34,2	31	3450	30,1	29,2
4	3480	31,3	31,0	18	2620	26,2	25,7	32	3600	31,4	31,4
5	3810	31,5	30,9	19	2900	26,7	26,4	33	2850	26,7	25,9
6	2330	24,5	23,9	20	1670	21,1	20,0	34	1590	22,1	21,4
7	1800	19,9	19,2	21	2540	24,1	23,9	35	3770	30,3	29,8
8	3110	27,3	27,2	22	3840	30,7	30,7	36	3850	32,0	30,6
9	3160	27,1	26,3	23	3800	32,7	32,6	37	2480	23,2	22,6
10	2310	24,0	23,9	24	4600	32,6	32,5	38	3570	30,3	30,3
11	4360	33,8	33,2	25	1900	22,1	20,8	39	2620	29,9	23,8
12	1880	21,5	21,0	26	2530	25,3	23,1	40	1890	20,8	18,4
13	3670	32,2	29,0	27	2920	30,8	29,8	41	3030	33,2	29,4
14	1740	22,5	22,0	28	4990	38,9	38,1	42	3030	28,2	28,2

Table 1: *Maximum compressive strength parallel to the grain (Y), density (X) and resin-adjusted density (Z) for 42 specimens of radiata pine*

The two competing models are

$$\begin{aligned} \mathcal{M}_1: &\qquad y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \nu^2) \\ \mathcal{M}_2: &\qquad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \tau^2), \end{aligned}$$

with the following prior distributions: $N(3000,\ 10^6)$ for both α and γ, $N(185, 10^4)$ for both β and δ and $IG(3, 1/(2\cdot 300^2))$ as for both σ^2 and τ^2, where $IG(a, b)$ denotes the *inverse gamma* distribution, whose density is

$$f(x) = \frac{1}{\exp[1/(bx)]\cdot[\Gamma(a)b^a x^{a+1}]}.$$

In this setting we compared our approach, denoted by RJ-MW, with that of Green (1995), denoted by RJ, that of Chib and Jelaizkov (2001), denoted by M, and that Mira and Nicholls (2000), denoted by M-MW. For the RJ algorithm we used 30,000 iterations, of which the first 5,000 are treated as burn-in, and two types of moves: within-model and across-model, each with probability 1/2. As in Han and Carlin (2001), we used the following proposal distributions to update the parameters within $\mathcal{M}_1$:

$$\alpha^* \sim N(\alpha, 5000), \quad \beta^* \sim N(\beta, 250), \quad \nu^{2*} \sim LN(\log(\nu^2), 1), \tag{14}$$

where $LN(\mu, \sigma^2)$ denotes the *log-normal* distribution. For the parameters γ, δ and τ^2 of $\mathcal{M}_2$ we used, respectively, the same proposals as for α, β and ν^2, while, to jump from $\mathcal{M}_1$ to $\mathcal{M}_2$, we simply let $(\alpha, \beta, \nu^2) = (\gamma, \delta, \tau^2)$; similarly to jump from $\mathcal{M}_2$ to $\mathcal{M}_1$. For the M method, based only on moves of type (14), we chosen any $\bar{\boldsymbol{\theta}}_k$ (in the notation of sec. 2.2) as the ML estimate of the corresponding parameter and, for any model, we used $N_1 = 2,000$ iterations, after a burn-in of 1,000, to compute the numerator and $N_2 = 2,000$ to compute the denominator; in the complex these are 10,000 iterations that roughly requires the same computing time of the 30,000 RJ iterations. Finally, for our algorithm we used 20,000 iterations in the whole, whose 2,000 treated as burn-in, so that approximately the same computing time of the other two algorithms is required.

The three algorithms were compared in term of efficiency in estimating the BF between $\mathcal{M}_2$ and $\mathcal{M}_1$, whose true value is $B_{21} = 4,862$ (see Green and O'Hagan, 1998), which clearly indicates that $\mathcal{M}_2$ has to be preferred to $\mathcal{M}_1$. The results of this comparison, based on 500 simulations, are in Table 2 that shows the following quantities for any algorithm:

- Mean $= \frac{1}{500}\sum_{i=1}^{500}\hat{B}_{21,i}$
- Standard error $= \sqrt{\frac{1}{500}\sum_{i=1}^{500}(\hat{B}_{21,i} - \text{Mean})^2}$
- Relative error $= \frac{1}{B_{21}}\sqrt{\frac{1}{500}\sum_{i=1}^{500}(\hat{B}_{21,i} - B_{21})^2}$,

where $\hat{B}_{21,i}$ is the estimate of B_{21} at the i-th trial of the simulation.

	RJ	M	M-MW	RJ-MW
Mean	7754.3	4887.7	4855.1	4862.3
Standard error	13,010.8	474.3	283.9	261.0
Relative error	274.13%	9.77%	5.84%	5.27%

Table 2: *Comparison of the algorithms for computing the Bayes factor for the data in Table 1.*

According to the previous results, the proposed algorithm is the most accurate in estimating B_{21}, whereas the RJ algorithm seems to the be the worst; this depends on the fact that $\mathcal{M}_2$ is much more likely than $\mathcal{M}_1$ and, therefore, the latter seldom jumps from $\mathcal{M}_1$ to $\mathcal{M}_2$. Han and Carlin (2001) overcome this problem by letting the priors of the models equal to $p(1) = 0.9995$ and $p(2) = 0.0005$; choosing these values, however, requires extra programming and computing time.

4.2 Logistic regression analysis

Dellaportas *et al.* (2001) compared several methods for selecting a hierarchical logistic regression model for the data in Table 3, concerning the relationship between the number of survivals, the patient condition (A) and the received treatment (B); these data are taken from Haely (1988).

Since there are two factors, we have 5 possible models: $\mathcal{M}_1$ (intercept); $\mathcal{M}_2$ (intercept+A); $\mathcal{M}_3$ (intercept+B); $\mathcal{M}_4$ (intercept+$A+B$); $\mathcal{M}_5$ (intercept+$A+B+A.B$). In particular, the full model

Patient condition	Antitoxin	Death	Survivals
Less severe	No	7	15
	Yes	5	15
More severe	No	22	4
	Yes	15	6

Table 3: *Number of survivals classified according to patient condition (A) and received treatment (B)*

($\mathcal{M}_5$) is formulated as

$$Y_{ij} \sim Bin(n_{ij}, p_{ij}), \qquad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB}, \tag{15}$$

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 surviving for the patients with condition i who received treatment j. Dellaportas *et al.* (2001) 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 independent. The same assumptions are made for any reduced model. Finally, the following proposal was used to jointly update the parameters within the same model (within-model move)

$$(\mu, \mu_2^A, \mu_2^B, \mu_{22}^{AB})' \sim N((-0.47, -0.87, 0.56, -0.17)', \mathrm{diag}(0.27, 0.27, 0.28, 0.27))$$

and also to jump from a model to another (across-model move).

Also in this setting we compared the proposed approach with those of Green (1995), Chib and Jelaizkov (2001) and Mira and Nicholls (2000). For the RJ algorithm we used 20,000 iterations, discarding the first 4,000 as burn-in, and only across-model moves. For the M method, we used $N_1 = 1,500$ iterations, after a burn-in of 500, for computing the numerator and $N_2 = 1,000$ for computing the denominator, in the complex 12,500 iterations. Finally, for our algorithm we used in the whole 16,000 iterations, with the first 4,000 treated as burn-in, chosen so that the computing time required by our algorithm is roughly the same of the other two. As in Section 4.1, the algorithms have been compared, from the point of view of the efficiency in estimating the BF, on the basis of 500 Monte Carlo simulations; the results of this comparison are shown in the Table 4 (as true value of the Bayes factor we took the overall means).

The RJ algorithm seems, again, to be the least efficient in estimating the BFs, even if the loss in terms of efficiency to the other methods is not so dramatic as is Section 4.1. Our algorithm performs generally better that the others. In particular, it has the smallest relative error in estimating B_{21}, B_{43} and B_{54}. Instead, the M-MW method has the smallest relative error in estimating B_{32}, but our method is almost as efficient.

		B_{21}	B_{32}	B_{43}	B_{54}
RJ	Mean	101.3580	0.0231	38.4289	0.1179
	Standard error	12.8320	0.0031	4.6844	0.0040
	Relative error	12.90%	13.38%	12.15%	3.43%
M	Mean	99.8230	0.0228	39.0250	0.1179
	Standard error	1.1829	0.0005	0.8505	0.0026
	Relative error	1.19%	2.08%	2.18%	2.20%
M-MW	Mean	99.8250	0.0228	39.0440	0.1179
	Standard error	1.0333	0.0004	0.7122	0.0021
	Relative error	1.04%	1.76%	1.82%	1.78%
RJ-MW	Mean	99.7480	0.0228	39.0500	0.1178
	Standard error	0.8641	0.0004	0.6990	0.0017
	Relative error	0.87%	1.85%	1.79%	1.43%

Table 4: *Comparison of the algorithms for computing the Bayes factor for the data in Table 3*

References

Bartolucci, F. and Scaccia, L. (2003), A new approach for estimating the BF, *Thechincal Report* University of Perugia, available at: www.stat.unipg.it/∼bart

Brooks, S.P., Giudici, P. and Roberts, G.O. (2003), Efficient Construction of Reversible Jump MCMC Proposal Distributions (with discussion), *Journal of the Royal Statistical Society, Series B*, **65**, p. 3-55.

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

Chen, M. H. and Shao, Q. M. (1997), On Monte Carlo methods for estimating ratios of normalizing constants, *The Annals of Statistics*, **25**, pp. 1563-1594.

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

Chib, S. (1995), Marginal output form the Gibbs output, *Journal of the American Statistical Association*, **90**, pp. 1313-1321.

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

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2001), On Bayesian model and variable selection using MCMC, *Statistics and Computing*, **12**, pp. 27-36.

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

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

Green, P. J. (2003), Trans-dimensional Markov chain Monte Carlo, in press.

Green, P. J. and O'Hagan, A. (1998), Carlin and Chib do not need to sample from pseudo-priors, *Research Report 98-1*, Department of Statistics, University of Nottingham.

Han, C. and Carlin, B. P. (2001), MCMC methods for computing Bayes factors: a comparative review, *Journal of the American Statistical Association*, **96**, pp. 1122-1132.

Jeffreys, H. (1935), Some Tests of Significance, Treated by Theory of Probability, *Proceeding of the Cambridge Philosophycal Society*, **31**, pp. 203-222.

Jeffreys, H. (1961), *Theory of Probability, 3rd ed.*, Oxford University press.

Kass, R. E. and Raftery, A. E. (1995), Bayes factors, *Journal of the America Statistical Association*, **90**, pp. 773-795.

Lavine, M. and Schervish, M. J. (1999), Bayes factors: what they are and what they are not, *The American Statistician*, **53**, pp. 119-122.

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

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

A. Mira and G. Nicholls (2000), Bridge estimation of the probability density at a point, Statistica Sinica, to appear. Mathematics Department, University of Auckland, *Technical Report* n. 456.

A. Mira e G. Nicholls, Bridge estimation of the probability density at a point, *Statistica Sinica*, to appear. Quaderno di Ricerca numero 2001/7, Dipartimento di Economia, Università dell'Insubria

Richardson, S. and Green, P. J. (1997), reply to the discussion of the paper On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society, ser. B*, **59**, pp. 784-790.

Williams, E. (1959), *Regression Analysis*, New York: Wiley.