On some difficulties with a posterior probability approximation technique

Christian Robert, Jean-Michel Marin

On some difficulties with a posterior probability approximation technique

Christian Robert[1,2,4] and Jean-Michel Marin[2,3,4]
[1]CEREMADE, Université Paris Dauphine,
[2]CREST, INSEE, Paris,
and [3]INRIA Saclay Ile-de-France, Projet select, Université Paris-Sud

January 23, 2008

Abstract

In Scott (2002) and Congdon (2006, 2007), a new method is advanced to compute posterior probabilities of models under consideration. It is based solely on MCMC outputs restricted to single models, i.e., it is bypassing reversible jump and other model exploration techniques. While it is indeed possible to approximate posterior probabilities based solely on MCMC outputs from single models, as demonstrated by Gelfand and Dey (1994) and Bartolucci et al. (2006), we show that the proposals of Scott (2002) and Congdon (2006, 2007) are biased and advance several arguments towards this thesis, the primary one being the confusion between model-based posteriors and joint pseudo-posteriors.
Keywords: Bayesian model choice, posterior approximation, reversible jump, Markov Chain Monte Carlo (MCMC), pseudo-priors, unbiasedness, improperty.

1 Introduction

Model selection is a fundamental statistical issue and a clear asset of the Bayesian methodology but it faces severe computational difficulties because of the requirement to explore simultaneously the parameter spaces of all models under comparison with enough of an accuracy to provide sufficient approximations to the posterior probabilities of all models. When Green (1995) introduced reversible jump techniques, it was perceived by the community as the second MCMC revolution in that it allowed for a valid and efficient exploration of the collection of models and the subsequent literature on the topic exploiting reversible jump MCMC is a testimony to the appeal of this method. Nonetheless, the implementation of reversible jump techniques in complex situations may face difficulties or at least inefficiencies of its own and, despite some recent advances in the devising of the jumps underlying reversible jump MCMC (Brooks et al., 2003), the care required in the construction of those jumps often acts as a deterrent from its applications.

There are practical alternatives to reversible jump MCMC when the number of models under consideration is small enough to allow for a complete exploration of those models. Integral approximations using importance sampling techniques like those found in Gelfand and Dey (1994), based

[4]xian@ceremade.dauphine.fr and jean-michel.marin@inria.fr

on an harmonic mean representation of the marginal densities, and in Gelman and Meng (1998), focussing on the optimised selection of the importance function, are advocated as potential solutions, see Chen et al. (2000) for a detailed entry. The reassessment of those methods by Bartolucci et al. (2006) showed the connection between a virtual reversible jump MCMC and importance sampling (see also Chopin and Robert, 2007). In particular, those papers demonstrated that the output of MCMC samplers on each single model could be used to produce approximations of posterior probabilities of those models, via some importance sampling methodologies also related to Newton and Raftery (1994).

In Scott (2002) and Congdon (2006), a new and straightforward method is advanced to compute posterior probabilities of models under scrutinity based solely on MCMC outputs restricted to single models. While this simplicity is quite appealing for the approximation of those probabilities, we believe that both proposals of Scott (2002) and Congdon (2006) are inherently biased and we advance in this note several arguments towards this thesis. In addition, we notice that, to overcome the bias we thus exhibited, a valid solution would call for the joint simulation of parameters under all models (using priors or pseudo-priors), unless the alternative advanced by Green and O'Hagan (1998) is is used, and this step would thus loose the primary appeal of the methods against the one proposed by Carlin and Chib (1995), from which both Scott (2002) and Congdon (2006) are inspired.

2 The methods

In a Bayesian framework of model comparison (see, e.g., Robert, 2001), given D models in competition, $\mathfrak{M}_k$, with densities $f_k(y|\theta_k)$, and prior probabilities $\varrho_k = P(M = k)$ $(k = 1, \ldots, D)$, the posterior probabilities of the models $\mathfrak{M}_k$ conditional on the data y are given by

$$P(M = k|y) \propto \varrho_k \int f_k(y|\theta_k)\pi_k(\theta_k)\,\mathrm{d}\theta_k\,,$$

the proportionality term being given by the sum of the above and M denoting the unknown model index.

In the specific setup of hidden Markov models, the solution of Scott (2002, Section 4.1) is to generate simultaneously and independently D MCMC chains

$$(\theta_k^{(t)})_t\,, \qquad 1 \le k \le D\,,$$

with stationary distributions $\pi_k(\theta_k|y)$ and to approximate $P(M = k|y)$ by

$$\tilde{\varrho}_k(y) \propto \varrho_k \sum_{t=1}^{T} f_k(y|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j\, f_j(y|\theta_j^{(t)})\,,$$

as reported in formula (21) of Scott (2002), with the mention that *"(21) averages the D likelihoods corresponding to each θ_j over the life of the Gibbs sampler" (p.347),* the later being understood as *"independently sampled D parallel Gibbs samplers" (p.347).*

From a more general perspective, the proposal of Congdon (2006) for an approximation of the $P(M = k|y)$'s follows both from Scott's (2002) approximation and from the pseudo-prior construction of Carlin and Chib (1995) that predated reversible jump MCMC by saturating the

parameter space with an artificial simulation of all parameters at each iteration. However, due to a very special (and, we believe, mistaken) choice of pseudo-priors discussed below, Congdon's (2006, p.349) approximation of $P(M = k|y)$ eventually reduces to the estimator

$$\hat{\varrho}_k(y) \propto \varrho_k \sum_{t=1}^{T} f_k(y|\theta_k^{(t)})\pi_k(\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)})\pi_j(\theta_j^{(t)}) \,,$$

where the $\theta_k^{(t)}$'s are samples from $\pi_k(\theta_k|y)$ (or approximate samples obtained by an MCMC algorithm). (A very similar proposal is found in Congdon (2007) and, while some issues like the use of improper pseudo-priors discussed in Section 3.3 are corrected, the fundamental difficulty of simulating from the wrong target remains. We thus chose to address primarily the initial paper by Congdon (2006), especially since it generated follow-up papers like Chen et al. (2008) and since its inherent simplicity is likely to appeal to unwary readers.)

Although both approximations $\tilde{\varrho}_k(y)$ and $\hat{\varrho}_k(y)$ differ in their expressions, they fundamentally relate to the same notion that parameters from other models can be ignored when conditioning on the model index M. This approach is therefore bypassing the simultaneous exploration of several parameters spaces and restricts the simulation to marginal samplers on each separate model. This feature is very appealing since it cuts most of the complexity from the schemes both of Carlin and Chib (1995) and of Green (1995). We however question the foundations of those approximations as presented in both Scott (2002) and Congdon (2006, 2007) and advance below arguments that both authors are using incompatible versions of joint distributions on the collection of parameters that jeopardise the validity of the approximations.

3 Difficulties

The sections below expose the difficulties found with both methods, following the points made in Scott (2002) and Congdon (2006), respectively. The fundamental difficulty with their approaches appears to us to be related to a confusion between the model dependent simulations and the joint simulations based on a pseudo-prior scheme as in Carlin and Chib (1995). Once this difficulty is resolved, it appears that the approximation of $P(M = k|y)$ by $\hat{P}(M = k|y)$ does require a joint simulation of all parameters and thus that the solutions proposed in Scott (2002) and Congdon (2006) are of the same complexity as the proposal of Carlin and Chib (1995).

3.1 Incorrect marginals

We denote by $\theta = (\theta_1, \ldots, \theta_D)$ the collection of parameters for all models under consideration. Both Scott (2002) and Congdon (2006) start from the representation

$$P(M = k|y) = \int P(M = k|y, \theta)\pi(\theta|y) \, \mathrm{d}\theta$$

to justify the approximation

$$\hat{P}(M = k|y) = \sum_{t=1}^{T} P(M = k|y, \theta^{(t)})/T \,.$$

This is indeed an unbiased estimator of $P(M = k|y)$ provided the $\theta^{(t)}$'s are generated from the correct (marginal) posterior

$$\pi(\theta|y) = \sum_{k=1}^{D} P(\theta, M = k|y) \tag{1}$$

$$\propto \sum_{k=1}^{D} \varrho_k \, f_k(y|\theta_k) \prod_j \pi_j(\theta_j)$$

$$= \sum_{k=1}^{D} \varrho_k \, m_k(y) \, \pi_k(\theta_k|y) \prod_{j \neq k} \pi_j(\theta_j) \, . \tag{2}$$

In both papers, the $\theta^{(t)}$'s are instead simulated as independent outputs from the componentwise posteriors $\pi_k(\theta_k|y)$ and this divergence jeopardises the validity of the approximation. The error in their interpretations stems from the fact that, while the $\theta_k^{(t)}$'s are (correctly) independent given the model index M, this independence does not hold once M is integrated out, which is the case in the above approximation $\hat{P}(M = k|y)$.

3.2 MCMC versus marginal MCMC

When Congdon (2006) defines a Markov chain $(\theta^{(t)})$ at the top of page 349, he indicates that the components of $\theta^{(t)}$ are made of independent Markov chains $(\theta_k^{(t)})$ simulated with MCMC samplers related to the respective marginal posteriors $\pi_k(\theta_k|y)$, following the approach of Scott (2002). The aggregated chain $(\theta^{(t)})$ is thus stationary against the product of those marginals,

$$\prod_{k=1}^{D} \pi_k(\theta_k|y) \, .$$

However, in the derivation of Carlin and Chib (1995), the model is defined in terms of (1) and the Markov chain should thus be constructed against (1), not against the product of the model marginals. Obviously, in the case of Congdon (2006), the fact that the pseudo-joint distribution does not exist because of the flat prior assumption (see Section 3.3) prevents this construction but, in the event the flat prior is replaced with a proper (pseudo-) prior (as in Congdon, 2007), the same statement holds: the probabilistic derivation of $P(M = k|y)$ relies on the pseudo-prior construction and, to be valid, it does require the completion step at the core of Carlin and Chib (1995), where parameters *need* to be simulated from the pseudo-priors.

Similarly, in Scott (2002), the target of the Markov chain $(\theta^{(t)}, M^{(t)})$ should be the distribution

$$P(\theta, M = k|y) \propto \pi_k(\theta_k) \, \varrho_k \, f_k(y|\theta_k) \prod_{j \neq k} \pi_j(\theta_j)$$

and the $\theta_j^{(t)}$'s should thus be generated from the prior $\pi_j(\theta_j)$ when $M^{(t)} \neq j$—or equivalently from the corresponding marginal if one does not condition on $M^{(t)}$, but simulating a Markov chain with stationary distribution (2) is certainly a challenge in many settings if the latent variable decomposing the sum is not to be used.

Since, in both Scott (2002) and Congdon (2006), the $(\theta^{(t)})$'s are not simulated against the correct target, the resulting averages of $P(M = k|y, \theta^{(t)})$, $\tilde{\varrho}_k(y)$ and $\hat{\varrho}_k(y)$, will both be biased, as demonstrated in the example of Section 3.4.

3.3 Improperty of the posterior

When resorting to the construction of pseudo-posteriors adopted by Carlin and Chib (1995), Congdon (2006) uses a *flat prior* as pseudo-prior on the parameters that are not in model $\mathfrak{M}_k$. More precisely, the joint prior distribution on (θ, M) is given by Congdon's (2006) formula (2),

$$\begin{aligned} P(\theta, M = k) &= \pi_k(\theta_k)\, \varrho_k \prod_{j \neq k} \pi(\theta_j | M = k) \\ &= \pi_k(\theta_k)\, \varrho_k \,, \end{aligned}$$

which is indeed equivalent to assuming a flat prior as pseudo-prior on the parameters θ_j that are not in model $\mathfrak{M}_k$.

Unfortunately, this simplifying assumption has a dramatic consequence in that the corresponding joint posterior distribution of θ is never defined (as a probability distribution) since

$$\pi(\theta|y) = \sum_{k=1}^{D} \pi_k(\theta_k|y)\, P(M = k|y)$$

does not integrate to a finite value in any of the θ_k's (unless their support is compact). When Congdon (2006) points out *"that it is not essential that the priors for $P(\theta_{j\neq k}|M = k)$ are improper" (p.348),* the whole issue is that they *cannot* be improper. (An alternative is to implement the jumping scheme of Green and O'Hagan (1998), in which case pseudo-priors become irrelevant, but the estimator of $P(\theta, M = k)$ then reduces to Scott's (2002) $\tilde{\varrho}_k$.)

The fact that the posterior distribution on the saturated vector $\theta = (\theta_1, \ldots, \theta_D)$ does not exist obviously has dire consequences on the subsequent derivations, since a positive recurrent Markov chain with stationary distribution $\pi(\theta|y)$ cannot be constructed. Similarly, the fact that

$$P(M = k|y) = \int P(\theta, M = k|Y)\, \mathrm{d}\theta$$

does not hold any longer.

Note that Scott (2002) does not follow the same track: when defining the pseudo-priors in his formula (20), he uses the product definition[1]

$$P(\theta, M = k) = \pi_k(\theta_k)\, \varrho_k \prod_{j \neq k} \pi_j(\theta_j) \,,$$

which (seemingly) means that the true priors are also used as pseudo-priors across all models. However, we stress that Scott (2002) does not refer to the construction of Carlin and Chib (1995) in his proposal.

3.4 Illustration

We now proceed through a toy example where all posterior quantities can be computed in order to evaluate the bias brought by both approximations.

[1]The indices on the priors have been added to make notations coherent, since Scott (2002) denotes all priors with the same letter p.

Example 1. Consider the case when a model $\mathfrak{M}_1 : y|\theta \sim \mathcal{U}(0,\theta)$ with a prior $\theta \sim \mathcal{E}xp(1)$ is opposed to a model $\mathfrak{M}_2 : y|\theta \sim \mathcal{E}xp(\theta)$ with a prior $\theta \sim \mathcal{E}xp(1)$. We also assume equal prior weights on both models: $\varrho_1 = \varrho_2 = 0.5$.

The marginals are then

$$m_1(y) = \int_y^\infty \theta^{-1} e^{-\theta}\, \mathrm{d}\theta = \mathrm{E}_1(y)\,,$$

where E_1 denotes the exponential integral function tabulated both in Mathematica and in the GSL library, and

$$m_2(y) = \int_0^\infty \theta e^{-\theta(y+1)}\, \mathrm{d}\theta = \frac{1}{(1+y)^2}\,.$$

For instance, when $y = 0.2$, the posterior probability of $\mathfrak{M}_1$ is thus equal to

$$\begin{aligned} P(M = 1|y) &= m_1(y)/\{m_1(y) + m_2(y)\} \\ &= \mathrm{E}_1(y)/\{\mathrm{E}_1(y) + (1+y)^{-2}\} \\ &\approx 0.6378\,, \end{aligned}$$

while, for $y = 0.9$, it is approximately 0.4843. This means that, in the former case, the Bayes factor of $\mathfrak{M}_1$ against $\mathfrak{M}_2$ is $B_{12} \approx 1.760$, while for the later, it decreases to $B_{12} \approx 0.939$.

The posterior on θ in model $\mathfrak{M}_2$ is a gamma $\mathcal{G}a(2, 1+y)$ distribution and it can thus be simulated directly. For model $\mathfrak{M}_1$, the posterior is proportional to $\theta^{-1}\exp(-\theta)$ for θ larger than y and it can be simulated using a standard accept-reject algorithm based on an exponential $\mathcal{E}xp(1)$ proposal translated by y.

Using simulations from the true (marginal) posteriors and the approximation of Congdon (2006), the numerical value of $\hat{\varrho}_1(y)$ based on 10^6 simulations is 0.7919 when $y = 0.2$ and 0.5633 when $y = 0.9$, which translates into Bayes factors of 3.805 and of 1.288, respectively. For the approximation of Scott (2002), the numerical value of $\tilde{\varrho}_1(y)$ is 0.6554 (corresponding to a Bayes factor of 1.898) when $y = 0.2$ and 0.6789 when $y = 0.9$ (corresponding to a Bayes factor of 2.11), based on the same simulations. Note that in the case $y = 0.9$, a selection based on either approximation of the Bayes factor would select the wrong model.

If we use instead a correct simulation from the joint posterior (2), which can be achieved by using a Gibbs scheme with target distribution $P(\theta, M = k|y)$, we then get a proper MCMC approximation to the posterior probabilities by the $\hat{P}(M = k|y)$'s. For instance, based on 10^6 simulations, the numerical value of $\hat{P}(M = 1|y)$ when $y = 0.2$ is 0.6370, while, for $y = 0.9$, it is 0.4843. Note that, due to the impropriety difficulty exposed in Section 3.3, the equivalent correction for Congdon's (2006) scheme cannot be implemented.

In Figure 1, the three approximations are compared to the exact value of $P(M = 1|y)$ for a range of values of y. The correct simulation produces a graph that is indistinguishable from the true probability, while Congdon's (2006) approximation stays within a reasonable range of the true value and Scott's (2002) surprisingly drifts apart for most values of y.

◀

The correspondence of what is essentially Carlin and Chib's (1995) scheme with the true numerical value of the posterior probabilities is obviously unsurprising in this toy example but more advanced setups see the approximation degenerate, since the simulations from the prior are most often inefficient, especially when the number of models under comparison is large. This is the reason why Carlin and Chib (1995) introduced pseudo-priors that were closer approximations to the true posteriors.

Acknowledgements

Both authors are grateful to Brad Carlin for helpful discussions and to Antonietta Mira for providing a perfect setting for this work during the ISBA-IMS "MCMC'ski 2" conference in Bormio, Italy. The second author is also grateful to Kerrie Mengersen for her invitation to "Spring Bayes 2007" in Coolangatta, Australia, that started our reassessment of those papers. This work had been supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2006-2008 project Adap'MC.

References

Bartolucci, F., Scaccia, L., and Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, 93:41–52.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Royal Statist. Soc. Series B*, 65(1):3–55.

Carlin, B. and Chib, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Roy. Statist. Soc. (Ser. B)*, 57(3):473–484.

Chen, C., Gerlach, R., and So, M. (2008). Bayesian model selection for heteroskedastic models. *Advances in Econometrics*, 23. To appear.

Chen, M., Shao, Q., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Chopin, N. and Robert, C. (2007). Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Technical Report 2007-46, CEREMADE, Université Paris Dauphine.

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Comput. Stat. Data Analysis*, 50:346–357.

Congdon, P. (2007). Model weights for model choice and averaging. *Statistical Methodology*, 4(2):143–157.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. (Ser. B)*, 56:501–514.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science*, 13:163–185.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. and O'Hagan, T. (1998). Model choice with mcmc on product spaces without using pseudo-priors. Technical report, University of Nottingham.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference by the weighted likelihood boostrap (with discussion). *J. Royal Statist. Soc. Series B*, 56:1–48.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st Century. *J. American Statist. Assoc.*, 97:337–351.

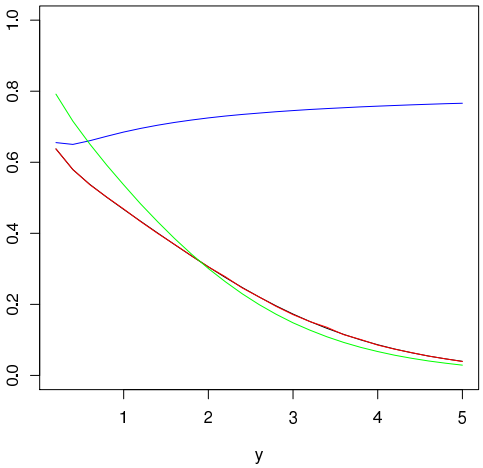

Figure 1: Comparison of three approximations of $P(M = 1|y)$ with the true value (in black): Scott's (2002) approximation (in blue), Congdon's (2006) approximation (in green), and correction of Scott's (2002) approximation (in red), indistinguishable from the true value (based on $N = 10^6$ simulations).