Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis 50 (2006) 346–357

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Bayesian model choice based on Monte Carlo estimates of posterior model probabilities

Peter Congdon*

Department of Geography, Queen Mary University of London, Mile End Road, London E1 4NS, UK

Received 13 July 2004; received in revised form 1 August 2004; accepted 1 August 2004
Available online 28 August 2004

Abstract

A range of approximate methods have been proposed for model choice based on Bayesian principles, given the problems involved in multiple integration in multi-parameter problems. Formal Bayesian model assessment is based on prior model probabilities $P(M = j)$ and posterior model probabilities $P(M = j|Y)$ after observing the data. An approach is outlined here that produces posterior model probabilities and hence Bayes factor estimates but not marginal likelihoods. It uses a Monte Carlo approximation based on independent MCMC sampling of two or more different models. While parallel sampling of the models is not necessary, such a form of sampling facilitates model averaging and assessing the impact of individual observations on the overall estimated Bayes factor. Three worked examples used before in model choice studies illustrate application of the method.

Keywords: Model choice; Posterior model probability; Bayes factor; Parallel sampling; Product space search

1. Introduction

A range of methods have been proposed for model choice and diagnosis based on Bayesian principles. For example, among the questions that regression model choice might include are choice among subsets of regressor variables, whether response and regressor variables need to be transformed, and whether a linear sum of regression effects should be used or various non-linear forms including general additive models. Formal Bayesian model assessment

* Tel.: +1-207-882-7760; fax: +1-208-981-6276.
E-mail address: p.congdon@qmul.ac.uk (P. Congdon).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.08.001

is based on prior model probabilities $\pi_k = P(M = k)$, where M is a multinomial model indicator, and posterior model probabilities

$$\rho_k = P(M = k|Y) = P(M = k)P(Y|M = k) \Bigg/ \sum_{j=1}^{K} [P(M = j)P(Y|M = j)] \, ,$$

where $P(Y|M = k)$ is the marginal likelihood for model k. Let θ_k be the parameter set associated with model k, then

$$P(M = k|Y) = P(M = k) \int P(Y|\theta_k)P(\theta_k)\,\mathrm{d}\theta_k \Bigg/ \sum_{j=1}^{K} \{P(M = j) \times \int P(Y|\theta_j)P(\theta_j)\,\mathrm{d}\theta_j\}.$$

The Bayes factor

$$B_{21} = P(Y|M = 2)/P(Y|M = 1) = [P(M = 2|Y)/P(M = 1|Y)]/[P(M_2)/P(M_1)]$$

is then used for comparing models. It has been suggested that values of $\log_e(B_{21})$ above three provide strong support for model 2, and those above five provide very strong support (Raftery, 1996).

Several approximations to the marginal likelihood or Bayes factor have been suggested. Gelfand and Dey (1994) recommend using an importance density approximation for the posterior $P(\theta_k|Y)$ while Newton and Raftery (1994) suggest the harmonic mean of the likelihoods, and propose methods of stabilising this estimator. Chib (1995) proposes a method for approximating marginal likelihoods based on the identity $P(Y|M=k)=P(Y|\theta_k)P(\theta_k|M=k)/P(\theta_k|Y)$. Path sampling (Gelman and Meng, 1998; Song and Lee, 2002) involves constructing a path to link two models being compared, and estimating the Bayes factor as a ratio of normalizing constants.

An alternative to ‘direct’ MCMC approaches which consider models singly (e.g. Chib, 1995) are model search methods which simulate over both model parameters and model indicators to produce samples from the posterior $p(k, \theta_k|Y)$. Examples are the reversible jump MCMC sampler (Green, 1995) that is developed on the basis of a detailed balance condition for $p(k, \theta_k|Y)$ and the Carlin–Chib (1995) method that involves a product space for parameters of all models under consideration and a model indexing variable M. The method of the current paper is also a product space method since it ‘keeps track of all θ_k not only the current one’ (Green, 2003). The approach here accumulates evidence on model relative performance without involving model jumps, but in a way consistent with a Bayesian approach. Wasserman (2000) and Kadane and Lazar (2004) note that the essential feature of Bayesian model choice and averaging is that it incorporates prior information on the parameters and this is true of the multiple model strategy proposed here. Links between this paper’s method and model search methods are considered in Section 3 and Appendix 2 below.

2. Prior assumptions

The approach proposed produces a Monte Carlo approximation to posterior model probabilities based on independent MCMC sampling of two (or more) different models. Following Carlin and Chib (1995), Godsill (2001) and Scott (2002, p. 347), the multinomial model vector $M \in (1, \ldots, K)$ indicates which θ_k is relevant in the likelihood for Y, and Y is independent of $\{\theta_{k \neq j}\}$ given $M = k$. Hence

$$P(Y|M = k, \theta) = P(Y|M = k, \theta_k). \tag{1}$$

Also assume independence among the θ_j given model M, as in Carlin and Chib (1995, p. 475). Hence

$$\begin{aligned} P(\theta, M = k) &= P(\theta|M = k)P(M = k) \\ &= [P(\theta_k|M = k)P(M = k)] \prod_{j \neq k}^{K} P(\theta_j|M = k). \end{aligned} \tag{2}$$

The posterior densities of θ_k are then conditionally independent given Y and $M = k$, and parallel sampling of K models yields the same posterior inferences as separate MCMC analyses of each model (Scott, 2002).

A further simplifying assumption is applied to the densities $P(\theta_{j \neq k}|M = k)$. It seems reasonable that each possible value of $\theta_{j \neq k}$ have equal weight when $M = k$, since specification of model k and model j is independent and model k cannot set a prior on parameters of a model $j \neq k$. This is expressed in the prior $P(\theta_{j \neq k}|M = k) = 1$, producing a simplification

$$\begin{aligned} P(\theta, M = k) &= [P(\theta_k|M = k)P(M = k)] \prod_{j \neq k}^{K} P(\theta_j|M = k) \\ &= P(\theta_k|M = k)P(M = k). \end{aligned} \tag{3}$$

Note that it is not essential that the priors for $P(\theta_{j \neq k}|M = k)$ are improper.

Parallel sampling is a computing convenience that assists in features such as model averaging. However, since the sampling chains of separate models are independent, one might run samples of the K models independently and then pool the relevant output. Typically $C_k > 1$ parallel chains are run for each model's parameters to assist in assessing convergence (Gelman et al., 1995); let $C_k = C$ for all k.

3. Monte Carlo approximation from parallel model sampling and links to model search algorithms

To assess the model performance, the goal is to estimate from the parallel or pooled samples of all models (i.e. from an MCMC output of dimension CK) the posterior probabilities

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

With samples $\{\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_k^{(t)}, \ldots, \theta_K^{(t)}), t = 1, T\}$ from $p(\theta_k|Y)$, $k = 1, \ldots, K$, a Monte Carlo estimate of $P(M = k|Y)$ is obtained (Scott, 2002) as

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

where (see Appendix 1)

$$P(M = k|Y, \theta^{(t)}) \propto P(Y|\theta_k^{(t)}, M = k)P(\theta_k^{(t)}|M = k)P(M = k) \tag{4}$$

and the normalising factor is the sum of the total model probabilities

$$G_j^{(t)} = P(Y, \theta_j^{(t)}, M = j) = P(Y|\theta_j^{(t)}, M = j)P(\theta_j^{(t)}|M = j)P(M = j).$$

For a default setting, such as equal prior model probabilities $P(M = k) = 1/K$, the model probabilities drop from this calculation, but for widely discrepant likelihoods unequal prior probabilities may be used (Carlin and Chib, 1995). The calculation in (4) is in practice based on finding the maximum $L_{\max}^{(t)}$ of the model log-likelihoods at each iteration

$$L_k^{(t)} = \log[P(Y|\theta_k^{(t)})P(\theta_k^{(t)}|M = k)P(M = k)]$$

and subtracting it from the other likelihoods, $D_k^{(t)} = L_k^{(t)} - L_{\max}^{(t)}$. Exponentiating $D_k^{(t)}$ then gives scaled total model likelihoods $G_k^{(t)}$, with

$$P(M = k|Y, \theta^{(t)}) = G_k^{(t)} \Bigg/ \sum_j G_j^{(t)} . \tag{5}$$

The averages w_k of the $w_k^{(t)} = P(M = k|Y, \theta^{(t)})$ over a long run may then be used to approximate Bayes factors as

$$B_{kj} = [w_k/w_j]/[P(M = k)/P(M = j)].$$

The above method generates $P(\theta_k|Y, k)$ by sampling from all K models and estimates $P(M = k|Y)$ from averages of $w_k^{(t)} = P(M = k|Y, \theta^{(t)})$. This is effectively equivalent to sampling from the joint parameter-indicator posterior since

$$P(k, \theta_k|Y) = P(M = k|Y)P(\theta_k|Y, k).$$

The outputs are therefore similar to those of MCMC methods for joint model-parameter search, such as reversible jump MCMC (Green, 1995) and the product space search of Carlin and Chib (1995). However, unlike these methods there is no mechanism proposing jumps between models. Rather model performance is monitored via total model probabilities $G_j^{(t)}$ for all models $j = 1, \ldots, K$ and all iterations $t = 1, \ldots, T$, and posterior model probabilities $P(M = k|Y)$ obtained by making the simplifying assumption $P(\theta_j|M = k, k \neq j) = 1$. This 'indifference prior' accords with a belief that model k is not meant to set priors on parameters of models $j \neq k$ and model k is indifferent to values taken by θ_j.

By contrast, model search methods may need 'informative' assumptions regarding pseudo-priors, particularly when they function as proposal densities or components of proposal densities. Appendix 2 considers reductions of the Carlin–Chib and RJMCMC algorithms when a jump is not proposed and shows them to be consistent with the approach of the current paper.

4. Model averaging and influence of individual observations

The model weights obtained at each iteration may also be used in model averaging. For example, consider a quantity $\Delta(\theta_k, Y)$ depending on both the data and the parameters of each model. The averaged sample of Δ at iteration *t* is

$$\Delta(\theta^{(t)}, Y) = \sum_k w_k^{(t)} \Delta(\theta_k^{(t)}, Y) \tag{6}$$

and the averaged quantity over all iterations is $\Delta = \sum_{t=1}^{T} \Delta(\theta^{(t)}, Y)/T$. Another option for model averaging is based on multinomial sampling of $\delta_k^{(t)}$ from the $w_k^{(t)}$ at each iteration and setting

$$\Delta(\theta^{(t)}, Y) = \sum_k \delta_k^{(t)} \Delta(\theta_k^{(t)}, Y). \tag{7}$$

One may also assess the role of individual observations in the overall Bayes Factor (Pettit and Young, 1990). As one possibility the model log-likelihood totals omitting the *i*th case likelihood, namely

$$A_{ik}^{(t)} = \log(H_{ik}^{(t)}) = \log\left[P(Y|\theta_k^{(t)})P(\theta_k^{(t)}|M=k)P(M=k)\right] - \log[P(Y_i|\theta_k^{(t)})],$$

are monitored and compared to the $A_{ij}^{(t)}$ for $j \neq k$, using a similar procedure to that above for the total model log-likelihoods $L_k^{(t)}$. So

$$w_{ik}^{(t)} = H_{ik}^{(t)} \bigg/ \sum_{j=1}^{K} H_{ij}^{(t)} ,$$

where $H_{ik}^{(t)}$ are rescaled versions of $H_{ik}^{(t)}$, and the $w_{ik}^{(t)}$ have averages w_{ik} over *T* iterations. The ratios

$$r_{ikj} = w_{ik}/w_{ij}$$

illustrate which observations, when omitted, most reduce or enhance the posterior probability of model *k* relative to other models. If case *i* contributed heavily to a Bayes factor favouring model *k* over model *j* then

$$b_{ikj} = r_{ikj}/[P(M=k)/P(M=j)] \tag{8}$$

will be lower than B_{kj}.

5. Illustrative examples

As a simple illustration suppose $n = 1000$ observations Y_i are generated from a $N(1, 1)$ density, and that two models are under consideration: M_1 is $Y_i \sim N(0, 1)$ and M_2 is $Y_i \sim N(\mu, 1)$ where $\mu \sim N(0, \tau)$, with $\tau = 100$. In this case 2 log B_{21} is known analytically, namely $(nY^{-2}/[1 + 1/\xi]) - \log(1 + \xi)$ where $\xi = n\tau$ (Raftery, 1996). The actual mean of the sampled data was $\bar{Y} = 0.9866$ (with $s^2 = 1.0021$) leading to an analytic value of 962 and the method returns a value of 965. Three more extensive worked examples based on data sets used in studies of Bayesian model choice are now considered (Winbugs code available at www.geog.qmul.ac.uk/staff/congdon.html).

5.1. Linear regression with alternative predictors: radiata pines data

Carlin and Chib (1995) and others have investigated prediction of Y, strength of radiata pines ($N = 42$) using either $X =$ density or $Z =$ adjusted density in linear regression models

$$M_1 \quad Y_i = \alpha_1 + \beta_1 X_i + \varepsilon_{i1},$$

$$M_2 \quad Y_i = \alpha_2 + \beta_2 Z_i + \varepsilon_{i2}$$

with $\varepsilon_{ik} \sim N(0, \tau_k)$. Let $\theta_k = (\alpha_k, \beta_k, \tau_k)$, $k = 1, 2$. For the parallel sampling approach, prior model probabilities $P(M_1) = 0.9999$ and $P(M_2) = 0.0001$ are assumed, with the priors $N(0, 10^8)$ priors on $\{\alpha_1, \alpha_2\}$, $N(185, 10^4)$ on $\{\beta_1, \beta_2\}$, and Ga(1,0.001) on $\{1/\tau_1, 1/\tau_2\}$. Convergence in the separate models is reached early (at under 1000 iterations) in a two chain run of 20,000 iterations for each model, with divergent starting values on θ_1 and θ_2. The default seed 314159 in Winbugs1.4 was used with Windows 2000 Professional. With $T = 19,000$, the average weights w_1 and w_2 are obtained as 0.6185 and 0.3815. See Fig. 1 for running quantile plot of w_2. The prior odds on model 2 are 1 to 9999 ($\approx$ 0.0001) and the posterior odds are $0.6168 = 0.3815/0.6185$. So model 2 is clearly favoured and the estimated Bayes factor B_{21} is 6167.5, with $\log(\mathrm{BF}_{21}) = 8.73$.

This estimate of the Bayes factor compares to 8.39 obtained by Carlin and Chib (1995) with priors $N(3000, 10^6)$ on $\{\alpha_1, \alpha_2\}$, $N(185, 10^4)$ on $\{\beta_1, \beta_2\}$, and $600^2\chi_6^{-2}$ priors on $\{\tau_1, \tau_2\}$. Song and Lee (2002) obtain a factor of 8.73 under the path sampling method, with priors $N(0, 10^6)$ on $\{\alpha_1, \alpha_2\}$, $N(0, 10^4)$ on $\{\beta_1, \beta_2\}$, and Ga(1,0.00333) on $\{1/\tau_1, 1/\tau_2\}$.

The ratios b_{i21} as in (8), which measure the impact of individual observations on the overall Bayes factor, vary from 928 to 11,900, with $\log(b_{i21})$ varying from 6.83 to 9.38. The lowest b_{i21} is for case 41, as in Table 1 in Carlin and Chib (1995), namely $\{Y, X, Z\} = (3030, 33.2, 29.4)$. This observation is better predicted under model 2 (posterior means 3976 and 3471 under models 1 and 2) so omitting it makes the Bayes factor deteriorate slightly. The highest b_{i21} is for case 40, $\{Y, X, Z\} = (1890, 20.8, 18.4)$, which is more markedly underpredicted under model 2 (posterior mean 1454) than under model 1 (with mean 1690). So omitting case 40 boosts the Bayes factor more towards model 2.

In an unpublished but often cited paper, Green and O’Hagan (2000) apply RJMCMC and analytic methods to comparing models 1 and 2, using the Carlin and Chibs priors. Also like Carlin and Chib they adopt prior probabilities $P(M_1) = 0.9995$ and $P(M_2) = 0.0005$. They obtain a posterior probability of 0.291 on model 1, so with prior odds of approximately

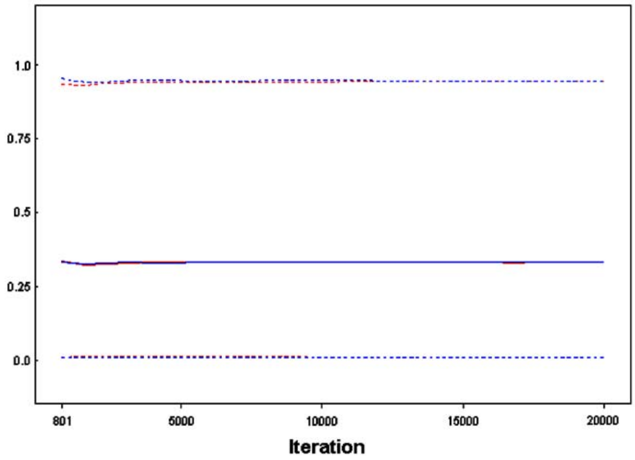

Fig. 1. Running quantile plot for w_2.

0.0005 on model 2 and posterior odds of 2.44 the Bayes factor favouring model 2 is 4870 (i.e. $\log B_{21} = 8.49$).

The method here was also applied with the Carlin and Chib priors and prior model probabilities. With seed 314159 and taking the last 19,000 of a two chain run of 20,000 iterations gives $w_1 = 0.243$, $w_2 = 0.757$. The Bayes factor estimate $B_{21} = 6231$ or $\log B_{21} = 8.74$ is very similar to the result with alternative priors as above.

5.2. Binomial regression

A second example involves the same data set considered by Dellaportas et al. (2002) and Green (2003), namely a 2 × 2 factorial study (patient frailty and treatment) with a binomial mortality response. All 5 possible models are considered (mean only, mean+frailty, mean+treatment, mean+frailty+treatment, and mean+frailty+treatment+frailty*treatment); priors are as in Dellaportas et al. The posterior model probabilities from the last 45,000 of a 50,000 iteration two chain run are (0.008, 0.507, 0.017, 0.405, 0.063). These are close to those reported by existing studies using RJMCMC or metropolised Carlin–Chib. Zero starting values were taken in one chain and starting values generated from the priors in the other, with seed 120967024.

5.3. Discrete mixture, normal components for the galaxy velocities

These data relate to $N = 82$ velocities of galaxies in the Corona Borealis region. Different analyses have suggested different numbers of components D in a discrete Normal mixture

$$f(Y|\mu_{jD}, \tau_{jD}, \pi_{jD}) = \sum_{j=1}^{D} \pi_{jD}\phi(Y_j|\mu_{jD}, \tau_{jD}).$$

Carlin and Chib (1995) compare $D = 3$ with $D = 4$ and adopt ‘graded priors’ on the μ_{jD}. For example, for the $D = 3$ model

$$\mu_{13} \sim N(9, 25),$$

$$\mu_{23} \sim N(18, 25)$$

and

$$\mu_{33} \sim N(30, 25).$$

Here models 1–3 with $D = 2, 3$, and 4, respectively, are compared with ‘ungraded’ priors $\mu_{jD} \sim N(20, 25)$, $j = 1, D$, Dirichlet priors for π_{jD} with D-vector components of 1, and $1/\tau_{jD}$ taken to be Ga(1,0.001).

Consider the estimated means for individual cases obtained at iteration t as $\mu_{C^{(t)}_{Di}}$, where $C^{(t)}_{Di}$ is a latent observation specific index variable with values between 1 and D. One might average the predicted means of selected observations (e.g. the minimum and maximum Y) to assess how well the different mixture models are representing extreme points. For example, with ranked data $Y_n \geqslant Y_{n-1} \geqslant \ldots Y_1$, the maximum Y value over models 1–3 with D ranging from 2 to 4, is predicted at iteration t as

$$\mu^{(t)}_n = w^{(t)}_1 \mu_{C^{(t)}_{2n}} + w^{(t)}_2 \mu_{C^{(t)}_{3n}} + w^{(t)}_3 \mu_{C^{(t)}_{4n}}.$$

A two chain run of 10,000 iterations (with 1000 burn in) and $P(M = k) = 1/3$ gives $w_1 = 0.0015$, $w_2 = 0.7876$, $w_3 = 0.2109$ so the $D = 3$ model is preferred though not overwhelmingly and different priors might affect the choice. This result is similar to Carlin and Chib (1995, p. 482) who compare $D=3$ and 4 using prior probabilities $P(D=3)=0.35$ and $P(D = 4) = 0.65$. They obtain $P(D = 3|Y)$ as 0.515, giving a Bayes factor of 1.75 in favour of $D = 3$. Moreno and Liseo (2003), using a default prior approach, compare $D=1, 2, 3, 4, 5$ and obtain $P(D=3|Y)=0.84$, $P(D=4|Y)=0.16$, with negligible weights on the other values of D. The maximum observed velocity (in thousands) is 34.28 and under models 1–3 as fitted here, this observation has mean 21.9, 31.65 and 31.8, respectively. Model 2 with $D=3$ thus adequately represents this point. The model averaged μ_n (with 95% credible interval) is 31.73 (29.5, 33.9) under the averaging in (6) and 31.73 (29.15,34.23) under (7).

6. Discussion

The method described here seeks to contribute to the development of sampling based Bayesian estimation of models. For large data sets, or comparing highly complex models, computing limitations may restrict analysis to small sets of models (e.g. $K = 2$). However, for small datasets with simple linear or discrete data regressions it may be possible to run a relatively large number of models in parallel. Such parallel sampling might be used in other ways: to allow for the impact of different priors on Bayes factors, one might average over a set R of ‘reasonable priors’ for a given model, or for $K > 1$, over both models and priors $r = 1, \ldots, R_k$.

As discussed in Section 3 and Appendix 2, the approach in the current paper is a special case of composite space MCMC sampling where jumps between models are not involved and so only evidence on model ratios $G_m^{(t)}/G_j^{(t)}$ needs to be accumulated. It is also noted that the indifference form of the pseudo-prior is compatible with the simplification of the composite space M–H algorithm proposed by Godsill (2001). There are well-known issues regarding the construction of proposal densities for RJMCMC and product space methods which are avoided in the simple alternative approach used here. There is also no problem in applying the method here to comparing complex models (e.g. with random effects where the model dimension is not known, inhibiting RJMCMC methods) or to comparison of non-nested models. Case study evaluation has shown posterior model probabilities close to those obtained with model search strategies.

This paper has considered formal model assessment through Bayes criteria but informal selection criteria (e.g. based on sampling new data $Y_{\text{new},k}$ from the different models at each iteration) may also be applied in concert with parallel sampling (Congdon, 2004). For example, as well as posterior predictive checks on one model (Gelman et al., 1995), one might assess whether one of two or more models 'checks better' against the data.

The WINBUGS code for the first two examples of the paper appear as an attachment to the multimedia files of the electronic version of this article in www.sciencedirect.com.

Appendix 1. Estimates of posterior model probabilities

By assumption

$$P(Y|M=k,\theta)=P(Y|M=k,\theta_k) \tag{A.1}$$

and under the 'indifference' pseudo-priors

$$P(\theta, M{=}k){=}[P(\theta_k|M{=}k)P(M{=}k)]\prod_{j\neq k}^{K}P(\theta_j|M{=}k){=}P(\theta_k|M{=}k)P(M{=}k). \tag{A.2}$$

A Monte Carlo estimate of $P(M=k|Y)$ is then obtained as

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

where

$$\begin{aligned} P(M=k|Y,\theta^{(t)}) &= P(M=k,Y,\theta^{(t)})/P(Y,\theta^{(t)}) \\ &= [P(Y|M=k,\theta^{(t)})P(\theta^{(t)},M=k)]/P(Y,\theta^{(t)}) \\ &= [P(Y|M=k,\theta_k^{(t)})P(\theta_k^{(t)}|M=k)P(M=k)]/P(Y,\theta^{(t)}), \end{aligned} \tag{A.3}$$

and the third line follows from (A.1)–(A.2). The denominator in (A.3) may be expressed

$$\begin{aligned} P(Y, \theta^{(t)}) &= \sum_{j=1}^{K} \{P(Y, \theta^{(t)}, M = j)\} \\ &= \sum_{j=1}^{K} \{P(Y|M = j, \theta^{(t)}) P(\theta^{(t)}|M = j) P(M = j)\}, \end{aligned}$$

and again from (A.1)–(A.2)

$$P(Y, \theta^{(t)}) = \sum_{j=1}^{K} P(Y|\theta_j^{(t)}, M = j) P(\theta_j^{(t)}|M = j) P(M = j).$$

Appendix 2. Simplification of product search and RJMCMC algorithms when a jump proposal is not made

Consider the Metropolis–Hastings version of the Carlin and Chib (1995) algorithm mentioned by Dellaportas et al. (2002). For simplicity let $M = k$ be denoted simply k. Then with current state (θ_k, k) a new model indicator m is proposed with probability J_{km}, and θ_m needs to be generated from the pseudo prior $P(\theta_m|k)$. The acceptance probability is then the minimum of 1 and

$$\begin{aligned} &[P(Y|\theta_m, m) P(\theta_m|m) P(m) J_{mk} \sum_{k \neq m}^{K} P(\theta_k|m)]/ \\ &\quad [P(Y|\theta_k, k) P(\theta_k|k) P(k) J_{km} \sum_{m \neq k}^{K} P(\theta_m|k)]. \end{aligned} \tag{B.1}$$

To ensure smooth transitions between models it is necessary to assume $P(\theta_m|k) \approx P(\theta_m|m, Y)$, namely that the pseudo-prior $P(\theta_m|k)$, now effectively a proposal density (Godsill, 2001), is close to or equal to the posterior density of θ_m. Note that if moves between models are not made but evidence is accumulated on the relative performance of all models, for example via total model probabilities $G_j^{(t)}$ as in the multiple modelling strategy used in the current paper, then simplifying assumptions may be made. Taking $J_{mk} = J_{km}$ and $P(\theta_k|m) = P(\theta_m|k) = 1$, the ratio in (B.1) above becomes $G_m^{(t)}/G_j^{(t)}$.

The basic form of the reversible jump MCMC algorithm (Green, 1995) can be seen as generalisation of the Metropolis–Hastings algorithm to include a model indicator. Moves from (k, θ_k) to (m, θ_m) are proposed according to a density $q(m, \theta_m|k, \theta_k)$ and the acceptance probability is the minimum of 1 and

$$[p(\theta_m, m|Y) q(k, \theta_k|m, \theta_m)/[p(\theta_k, k|Y) q(m, \theta_m|k, \theta_k)].$$

More specialised forms of RJMCMC apply in special applications, such as when models are nested (Godsill, 2001). Both product space and RJMCMC algorithms are special cases of a composite space M–H algorithm that considers moves from current state (k, θ) to a potential new state (m, θ^*) where $\theta = (\theta_1, \ldots, \theta_K)$ and $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$ (Godsill, 2001; Chen et al., 2000). With proposal density $q(m, \theta^*|k, \theta)$, the acceptance probability is the minimum of 1 and

$$[p(\theta^*, m|Y)q(k, \theta|m, \theta^*)/[p(\theta, k|Y)q(m, \theta^*|k, \theta)], \tag{B.2}$$

where

$$p(\theta, k|Y) \propto R(k, \theta|Y) = p(k)p(Y|\theta_k, k)p(\theta_k|k)p(\theta_{[k]}|\theta_k, k).$$

Here $\theta_{[k]} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K)$ is θ omitting θ_k, and $p(\theta_{[k]}|\theta_k, k)$ is a pseudo-prior. Similarly

$$p(\theta^*, m|Y) \propto R(m, \theta^*|Y) = p(m)p(Y|\theta_m^*, m)p(\theta_m^*|m)p(\theta_{[m]}^*|\theta_m^*, m),$$

where $p(\theta_{[m]}^*|\theta_m^*, m)$ is a pseudo-prior. Godsill (2001, Section 2.3.1) assumes proposal densities of the form

$$q(m, \theta^*|k, \theta) = J_{mk}q(\theta_m^*|\theta_k)p(\theta_{[m]}^*|\theta_m^*, m),$$

$$q(k, \theta|m, \theta^*) = J_{km}q(\theta_k|\theta_m^*)p(\theta_{[k]}|\theta_k, k),$$

namely that the components of the proposal densities governing unused parameters $\theta_{[m]}^*$ and $\theta_{[k]}$ are pseudo-priors. On this assumption the pseudo-priors cancel in (B.2) and the acceptance rate is the minimum of 1 and

$$\begin{aligned}&[J_{km}q(\theta_k|\theta_m^*)p(m)p(Y|\theta_m^*, m)p(\theta_m^*|m)]/\\ &\quad[J_{mk}q(\theta_m^*|\theta_k)p(k)p(Y|\theta_k, k)p(\theta_k|k)].\end{aligned} \tag{B.3}$$

Again if model jumps are not made but evidence is still accumulated on the relative performance of all models at each iteration, then q and J terms can be omitted and the ratio in (B.3) becomes $G_m^{(t)}/G_j^{(t)}$. Hence the method of the current paper is consistent with product space and RJMCMC methods if model jumps are not made. It may be noted that the simplifications leading to (B.3) can also be obtained by assuming

$$p(\theta_{[k]}|\theta_k, k) = p(\theta_{[k]}|k) = 1 \text{ and } p(\theta_{[m]}^*|\theta_m^*, m) = p(\theta_{[m]}^*|m) = 1.$$

References

Carlin, B., Chib, S., 1995. Bayesian model choice via Markov Chain Monte Carlo methods. J. Roy. Statist. Soc. Ser. B 57, 473–484.

Chen, M., Shao, Q., Ibrahim, J., 2000. Monte Carlo Methods in Bayesian Computation, Springer, New York.

Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc. 90, 1313–1321.

Congdon, P., 2004. Bayesian predictive model comparison via parallel sampling. Computational Statistics and Data Analysis, in press.

Dellaportas, P., Forster, J., Ntzoufras, I., 2002. On Bayesian model and variable selection using MCMC. Statist. Comput. 12, 27–36.

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

Gelman, A., Meng, X., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statist. Sci. 13, 163–185.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. Bayesian Data Analysis, Chapman & Hall, London.

Godsill, S., 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graphical Statist. 10, 230–248.

Green, P., 1995. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Green, P., 2003. Trans-dimensional Markov chain Monte Carlo. In: Green, P., Hjort, N., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press, Oxford, pp. 179–198.

Green, P., O'Hagan, A., 2000. Model choice with MCMC on product spaces without using pseudo priors. Technical Report, Department of Statistics, University of Nottingham.

Kadane, J., Lazar, N., 2004. Methods and criteria for model selection. J. Amer. Statist. Assoc. 99, 279–290.

Moreno, E., Liseo, B., 2003. A default Bayesian test for the number of components in a mixture. J. Statist. Plann. Inference 111, 129–142.

Newton, M., Raftery, A., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Roy. Statist. Soc. Ser. B 56, 3–48.

Pettit, L., Young, K., 1990. Measuring the effect of observations on Bayes factors. Biometrika 77, 455–466.

Raftery, A., 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83, 251–266.

Scott, S., 2002. Bayesian methods for hidden Markov models: recursive computing in the 21st century. J. Amer. Statist. Assoc. 97, 337–351.

Song, X., Lee, S., 2002. A Bayesian model selection method with applications. Computational Statistics and Data Analysis 40, 539–557.

Wasserman, L., 2000. Bayesian model selection and model averaging. J. Math. Psychol. 44, 92–107.