Statistics and Computing **12**: 27–36, 2002

On Bayesian model and variable selection using MCMC

PETROS DELLAPORTAS*, JONATHAN J. FORSTER† and IOANNIS NTZOUFRAS**

**Department of Statistics, Athens University of Economics and Business, Patission 76, 10434 Athens, Greece*
petros@aueb.gr
†*Department of Mathematics, University of Southampton, Highfield, Southampton SO17 1BJ, UK*
jjf@maths.soton.ac.uk
***Department of Statistics, Athens University of Economics and Business, Patission 76, 10434 Athens, Greece*
jbn@stat-athens.aueb.gr

Received September 1998 and accepted September 2000

Several MCMC methods have been proposed for estimating probabilities of models and associated 'model-averaged' posterior distributions in the presence of model uncertainty. We discuss, compare, develop and illustrate several of these methods, focussing on connections between them.

Keywords: Gibbs sampler, independence sampler, Metropolis–Hastings, reversible jump

1. Introduction

A Bayesian approach to model selection proceeds as follows. Suppose that the data $\boldsymbol{y}$ are considered to have been generated by a model m, one of a set M of competing models. Each model specifies the distribution of $\boldsymbol{Y}$, $f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m)$ apart from an unknown parameter vector $\boldsymbol{\beta}_m \in B_m$, where B_m is the set of all possible values for the coefficients of model m. If $f(m)$ is the prior probability of model m, then the posterior probability is given by

$$f(m \mid \boldsymbol{y}) = \frac{f(m) f(\boldsymbol{y} \mid m)}{\sum_{m \in M} f(m) f(\boldsymbol{y} \mid m)}, \quad m \in M \tag{1}$$

where $f(\boldsymbol{y} \mid m)$ is the marginal likelihood calculated using $f(\boldsymbol{y} \mid m) = \int f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m \mid m) d\boldsymbol{\beta}_m$ and $f(\boldsymbol{\beta}_m \mid m)$ is the conditional prior distribution of $\boldsymbol{\beta}_m$, the model parameters for model m.

This integral is only analytically tractable in certain restricted examples. A further problem is that the size of the set of possible models M may be so great that calculation or approximation of $f(\boldsymbol{y} \mid m)$ for all $m \in M$ becomes infeasible. Therefore MCMC methods which generate observations from the joint posterior distribution $f(m, \boldsymbol{\beta}_m \mid \boldsymbol{y})$ of $(m, \boldsymbol{\beta}_m)$ have recently become popular for estimating $f(m \mid \boldsymbol{y})$ and $f(\boldsymbol{\beta}_m \mid m, \boldsymbol{y})$. The natural parameter space for $(m, \boldsymbol{\beta}_m)$ is

$$B = \bigcup_{m \in M} \{m\} \times B_m.$$

We focus on a straightforward independence sampler, and on the methods of Green ('reversible jump' 1995) and Carlin and Chib (1995), and describe a connection between them.

We also consider 'variable selection' problems where the models under consideration can be naturally represented by a set of binary indicator variables so that $M \subseteq \{0, 1\}^p$, where p is the total possible number of variables. We introduce a modification of Carlin and Chib's method for variable selection problems, which can be more efficient in certain examples.

2. MCMC model selection methods

2.1. *Independence sampler*

The most straightforward MCMC approach from generating from the posterior distribution $f(m, \boldsymbol{\beta}_m \mid \boldsymbol{y})$ over B is a standard Metropolis–Hastings approach. Given the current value of $(m, \boldsymbol{\beta}_m)$, a proposal $(m', \boldsymbol{\beta}'_{m'})$ is generated from some proposal distribution over B. If the proposal distribution has density $q(m', \boldsymbol{\beta}'_{m'} \mid m, \boldsymbol{\beta}_m)$ with respect to the natural measure on B, then the proposal is accepted as the next observation of the chain

0960-3174

with the usual Metropolis–Hastings acceptance probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y} \mid m', \boldsymbol{\beta}'_{m'}) f(\boldsymbol{\beta}'_{m'} \mid m') f(m') q(m, \boldsymbol{\beta}_m \mid m', \boldsymbol{\beta}'_{m'})}{f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m \mid m) f(m) q(m', \boldsymbol{\beta}'_{m'} \mid m, \boldsymbol{\beta}_m)}\right). \quad (2)$$

In practice, the proposal is constructed (like the prior) as a proposal for model m', followed by a proposal for model parameters $\boldsymbol{\beta}'_{m'} \mid m'$, so $q(m', \boldsymbol{\beta}'_{m'} \mid m, \boldsymbol{\beta}_m) = q(m' \mid m, \boldsymbol{\beta}_m) q(\boldsymbol{\beta}'_{m'} \mid m', m, \boldsymbol{\beta}_m)$. This approach is considered by Gruet and Robert (1997) for mixture models.

The independence sampler (Tierney 1994) is a special case of this approach, which is straightforward to implement. For the independence sampler, the proposal can be represented as $q(m', \boldsymbol{\beta}'_{m'})$ and is not allowed to depend on the current values $(m, \boldsymbol{\beta}_m)$. This approach is used by Clyde and Desimone–Sasinowska (1998). The independence sampler works best if the proposal q is a reasonable approximation to the target posterior distribution f. In the current context, this is clearly going to be difficult to achieve, as we require not just a reasonable approximation to $f(m \mid \boldsymbol{y})$ but, perhaps more importantly, a reasonable approximation to $f(\boldsymbol{\beta}_m \mid m, \boldsymbol{y})$ for every m. If $|M|$, the number of models is small, it may be possible to construct an approximation to each $f(\boldsymbol{\beta}_m \mid m, \boldsymbol{y})$ based on a pilot MCMC run within the model. When there are many models, an alternative approach will be required. This is discussed further in Section 4.

2.2. *Reversible jump*

Generally, the independence sampler is unlikely to represent the best strategy (Roberts 1996). Therefore, it seems sensible to allow the proposal $(m', \boldsymbol{\beta}'_{m'})$ to depend on the current values $(m, \boldsymbol{\beta}_m)$. However, for parameter spaces such as B, it is desirable to allow this in the most flexible way possible. In particular, the natural way to allow the parameters $\boldsymbol{\beta}'_{m'}$ of a proposed model to depend on the parameters $\boldsymbol{\beta}_m$ of the current model, may suggest a proposal distribution with sample space of lower dimension than $B_{m'}$. For example, where model m' is nested in m then we may want to propose $\boldsymbol{\beta}'_{m'}$ as a deterministic function of $\boldsymbol{\beta}_m$.

The standard Metropolis–Hastings acceptance probability (2) cannot be applied to this kind of proposal, over a sample space of reduced dimension. However, Green (1995) developed reversible jump MCMC for exactly this situation. The reversible jump approach for generating from $f(m, \boldsymbol{\beta}_m \mid \boldsymbol{y})$ is based on creating a Markov chain which can 'jump' between models with parameter spaces of different dimension in a flexible way, while retaining detailed balance which ensures the correct limiting distribution provided the chain is irreducible and aperiodic.

Suppose that the current state of the Markov chain is $(m, \boldsymbol{\beta}_m)$, where $\boldsymbol{\beta}_m$ has dimension $d(\boldsymbol{\beta}_m)$, then one version of the procedure is as follows:

- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{u}$ (which can be of lower dimension than $\boldsymbol{\beta}_{m'}$) from a specified proposal density $q(\boldsymbol{u} \mid \boldsymbol{\beta}_m, m, m')$.
- Set $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u}') = g_{m,m'}(\boldsymbol{\beta}_m, \boldsymbol{u})$ where $g_{m,m'}$ is a specified invertible function. Hence $d(\boldsymbol{\beta}_m) + d(\boldsymbol{u}) = d(\boldsymbol{\beta}_{m'}) + d(\boldsymbol{u}')$. Note that $g_{m,m'} = g^{-1}_{m,m'}$.
- Accept the proposed move to model m' with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y} \mid m', \boldsymbol{\beta}'_{m'}) f(\boldsymbol{\beta}'_{m'} \mid m') f(m') j(m', m) q(\boldsymbol{u}' \mid \boldsymbol{\beta}_m, m', m)}{f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m \mid m) f(m) j(m, m') q(\boldsymbol{u} \mid \boldsymbol{\beta}_{m'}, m, m')} \times \left|\frac{\partial g(\boldsymbol{\beta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\beta}_m, \boldsymbol{u})}\right|\right). \quad (3)$$

There are many variations of simpler versions of reversible jump that can be applied in specific model determination problems. In particular, if all parameters of the proposed model are generated directly from a proposal distribution, then $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u}') = (\boldsymbol{u}, \boldsymbol{\beta}_m)$ with $d(\boldsymbol{\beta}_m) = d(\boldsymbol{u}')$ and $d(\boldsymbol{\beta}_{m'}) = d(\boldsymbol{u})$, and the jacobian term in (3) is one. The acceptance probabilities in (2) and (3) are equivalent, and we have an independence sampler. Therefore the independence sampler is a special case of reversible jump. With the same proposals, but where the function $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u}') = g_{m,m'}(\boldsymbol{u}, \boldsymbol{\beta}_m)$ is not the identity then we have a more general Metropolis-Hastings algorithm where $\boldsymbol{\beta}'_{m'}$ is allowed to depend on $\boldsymbol{\beta}_m$. If $m' = m$, then the move is a standard Metropolis-Hastings step.

However, the real flexibility of the reversible jump formulation is that it allows us to use proposal distributions of lower dimension than $d(\boldsymbol{\beta}'_{m'})$. For example, if model m is nested in m' then, as suggested above, there may be an extremely natural proposal distribution and transformation function $g_{m,m'}$ (may be the identity function) such that $d(\boldsymbol{u}') = 0$ and $\boldsymbol{\beta}'_{m'} = g_{m,m'}(\boldsymbol{\beta}_m, \boldsymbol{u})$. Therefore, when the reverse move is proposed, the model parameters are proposed deterministically. See, for example, Dellaportas and Forster (1999).

2.3. *Carlin and Chib's method*

Carlin and Chib (1995) proposed using a Gibbs sampler to generate from the posterior distribution $f(m, \boldsymbol{\beta}_m \mid \boldsymbol{y})$. In order to do this, it is required to consider a Markov chain of realisations of $(m, \boldsymbol{\beta}_k : k \in M)$, and extract the marginal sample of $(m, \boldsymbol{\beta}_m)$. The parameter space for $(m, \boldsymbol{\beta}_k : k \in M)$ is $M \times \prod_{m \in M} B_m$. Therefore, a prior distribution for $(m, \boldsymbol{\beta}_k : k \in M)$ is no longer completely specified by $f(m)$ and $f(\boldsymbol{\beta}_m \mid m)$, so Carlin and Chib proposed the use of pseudopriors or linking densities $f(\boldsymbol{\beta}_k \mid m \neq k), k \in M$.

The full posterior conditional distributions are

$$f(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \{\boldsymbol{\beta}_l : l \neq k\}, m) \propto \begin{cases} f(\boldsymbol{y} \mid \boldsymbol{\beta}_m, m) f(\boldsymbol{\beta}_m \mid m) & k = m \\ f(\boldsymbol{\beta}_k \mid k \neq m) & k \neq m \end{cases} \quad (4)$$

and

$$f(m \mid \{\boldsymbol{\beta}_k : k \in M\}, \boldsymbol{y}) = \frac{A_m}{\sum_{k \in M} A_k} \quad (5)$$

where

$$A_m = f(\boldsymbol{y} \mid \boldsymbol{\beta}_m, m) \prod_{l \in M} [f(\boldsymbol{\beta}_l \mid m)] f(m).$$

When $k = m$, β_k are generated from the conditional posterior distribution $f(\beta_m \mid \boldsymbol{y}, m)$ and when $k \neq m$ from the corresponding pseudoprior, $f(\beta_k \mid m)$. For $k \neq m$ this approach involves generating directly from the pseudopriors, and therefore is optimised when each pseudoprior density $f(\beta_k \mid m)$ is close to the corresponding conditional posterior density $f(\beta_k \mid k, \boldsymbol{y})$. This may be achieved by a common pseudoprior density $f(\beta_k \mid m)$ for all $m \in M \backslash \{k\}$, which we denote by $f(\beta_k \mid k \neq m)$. The model indicator m is generated as a discrete random variable.

The main drawback of this method is the unavoidable specification of, and generation from, many pseudoprior distributions. Carlin and Chib (1995) point out that, as the pseudopriors do not enter the marginal posterior distributions $f(m, \beta_m)$ of interest, they should be chosen to make the method efficient. However, generation from $|M| - 1$ pseudopriors at every cycle of the Gibbs sampler is still required, and this is computationally demanding. Green and O'Hagan (1998) show that it is not necessary to generate from the pseudopriors for the chain to have the correct limiting distribution, but that this modification is unlikely to be efficient.

2.4. *'Metropolised' Carlin and Chib*

The Gibbs sampler proposed by Carlin and Chib (1995) requires the calculation of all A_k in the denominator of (5). An alternative approach is a hybrid Gibbs/Metropolis strategy, where the 'model selection' step is not based on the full conditional, but on a proposal for a move to model m', followed by acceptance or rejection of this proposal. If the current state is model m and we propose model m' with probability $j(m, m')$, then the acceptance probability is given by

$$\begin{aligned} \alpha &= \min\left(1, \tfrac{A_{m'} j(m', m)}{A_m j(m, m')}\right) \\ &= \min\left(1, \frac{f(\boldsymbol{y} \mid \beta_{m'}, m') f(\beta_{m'} \mid m') f(\beta_m \mid m') f(m') j(m', m)}{f(\boldsymbol{y} \mid \beta_m, m) f(\beta_m \mid m) f(\beta_{m'} \mid m) f(m) j(m, m')}\right) \end{aligned} \quad (6)$$

as all other pseudopriors cancel.

Note that when we are in model m and we propose model m', we require only values of β_m and $\beta_{m'}$ to calculate α in (6). Furthermore, we are assuming that model m' is proposed with probability $j(m, m')$, independently of the values of any model parameters. Therefore if we reverse the order of sampling from j and the full conditional distributions for β_k in (4), there is no need to sample from any pseudopriors other than that for m'. The method now consists of the following steps:

- Propose a new model m' with probability $j(m, m')$.
- Generate β_m from the posterior $f(\beta_m \mid \boldsymbol{y}, m)$.
- Generate $\beta_{m'}$ from the pseudoprior $f(\beta_{m'} \mid m \neq m')$.
- Accept the proposed move to model m' with probability α given by (6).

It is straightforward to see that by a simple modification ('Metropolising' the model selection step), the model selection step of Carlin and Chib's method becomes equivalent to an independence sampler. The pseudopriors become proposals in the independence sampler, and only one pseudoprior is used at each iteration. This independence sampler is combined with a Gibbs sampler which updates the parameters of the current model at each iteration, although this is not strictly necessary. We use this hybrid approach in the examples of Section 5.

More generally, Besag (1997) and Godsill (1998) show that Metropolis–Hastings algorithms in the product space $M \times \prod_{m \in M} B_m$ can be thought of as equivalent algorithms in B, with acceptance probabilities of the form (2). All these approaches are less flexible than general reversible jump, as they all require a proposal distribution of full dimension $d(\beta'_{m'})$ when model m' is proposed. Reversible jump on the other hand allows the model parameters of the proposed model to depend on the parameters of the current model in a totally general way through the function $g_{m',m}$. In particular, the dimension of the proposal distribution may be much less than the dimension of the proposed model.

For a simple illustration of this, consider example 3.1 of Carlin and Chib (1995) where there are two possible models, the non-nested regression models $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, $i = 1, \ldots, n (m = 1)$ and $Y_i \sim N(\gamma + \delta z_i, \tau^2)$, $i = 1, \ldots, n (m = 2)$. When it is proposed to change model (from $m = 1$ to $m = 2$ say) the independence sampler requires a proposal for (γ, δ, τ) independent of the current values of (α, β, σ). (Carlin and Chib's approach requires an equivalent pseudoprior). A reversible jump strategy can make use of these current values in a deterministic way, and a plausible proposal is

$$\begin{pmatrix} \gamma \\ \delta \\ \tau \end{pmatrix} = \begin{pmatrix} 1 & -\frac{a}{b} & 0 \\ 0 & \frac{1}{b} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \sigma \end{pmatrix} \quad (7)$$

where a and b are the least squares estimates of a linear regression of z on x.

When a move from $m = 2$ to $m = 1$ is made, (α, β, σ) are proposed through the inverse of (7). As both these proposals are purely deterministic, the proposal densities q disappear from the numerator and denominator of (3). From (7) it can be seen that the Jacobian is $1/b$ for a move from $m = 1$ to $m = 2$ and b for the reverse move.

In addition to the proposed model change outlined above, a Gibbs sampler step to update the parameters of the current model was used at each iteration. For the data analysed by Carlin and Chib (1995), this reversible jump chain proved to be quite mobile. More than one in three proposed model switches were accepted. No pilot runs or 'training' were required.

2.5. *Using posterior distributions as proposals*

Suppose that, for each m, the posterior density $f(\beta_m \mid m, \boldsymbol{y})$ is available, including the normalising constant which is the marginal likelihood $f(\boldsymbol{y} \mid m)$. If this distribution is used as a pseudoprior then the acceptance probability in (6) is

given by

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y} \mid m', \boldsymbol{\beta}'_{m'}) f(\boldsymbol{\beta}'_{m'} \mid m') f(m') j(m', m) f(\boldsymbol{\beta}_m \mid m, \boldsymbol{y})}{f(\boldsymbol{y} \mid m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m \mid m) f(m) j(m, m') f(\boldsymbol{\beta}'_{m'} \mid m', \boldsymbol{y})}\right)$$

$$= \min\left(1, B_{m'm} \frac{f(m') j(m', m)}{f(m) j(m, m')}\right)$$

where $B_{m'm}$ is the Bayes factor of model m' against model m. In practice, we cannot usually calculate $B_{m'm}$. In the special case where models are decomposable graphical models, Madigan and York (1995) used exactly this approach, which they called MC^3. Here there is no need to generate the model parameters $\boldsymbol{\beta}_m$ as part of the Markov chain. These can be generated separately from the known posterior distributions $f(\boldsymbol{\beta}_m \mid m, \boldsymbol{y})$ if required.

3. Variable selection

Many statistical models may be represented naturally as $(s, \boldsymbol{\gamma}) \in S \times \{0, 1\}^p$, where the indicator vector $\boldsymbol{\gamma}$ represents which of the p possible sets of covariates are present in the model and s represents other structural properties of the model. For example, for a generalised linear model s may describe the distribution, link function and variance function, and the linear predictor may be written as

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{X}_i \boldsymbol{\beta}_i \tag{8}$$

where $\boldsymbol{X}_i$ is the design matrix and $\boldsymbol{\beta}_i$ the parameter vector related to the ith term. In the following, we restrict consideration to variable selection aspects assuming that s is known, or dealt with in another way and therefore we substitute $\boldsymbol{\gamma}$ for model indicator m. For example, we can apply reversible jump to variable selection by substituting $\boldsymbol{\gamma}$ for m in (3).

Note that in some cases, it is sensible to set $f(\gamma_i \mid \boldsymbol{\gamma}_{\setminus i}) = f(\gamma_i)$ (where the subscript $\setminus i$ denotes all elements of a vector except the ith), whereas in other cases (e.g. hierarchical or graphical log-linear models) it is required that $f(\gamma_i \mid \boldsymbol{\gamma}_{\setminus i})$ depends on $\boldsymbol{\gamma}_{\setminus i}$; see Chipman (1996).

3.1. *Gibbs variable selection*

Here we introduce a modification of Carlin and Chib's Gibbs sampler which is appropriate for variable selection, and can be implemented in Gibbs sampler software such as BUGS. Furthermore, it does not require unnecessary generation from pseudopriors.

We specify the prior for $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ as $f(\boldsymbol{\gamma}, \boldsymbol{\beta}) = f(\boldsymbol{\gamma}) f(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$. If we consider a partition of $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}})$ corresponding to those components of $\boldsymbol{\beta}$ which are included ($\gamma_i = 1$) or not included ($\gamma_i = 0$) in the model, then the prior $f(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ may be partitioned into model prior $f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})$ and pseudoprior $f(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$.

The full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{9}$$

$$f(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto f(\boldsymbol{\beta}_{\setminus \gamma} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{10}$$

and

$$\frac{f(\gamma_i = 1 \mid \boldsymbol{\gamma}_{\setminus i}, \boldsymbol{\beta}, \boldsymbol{y})}{f(\gamma_i = 0 \mid \boldsymbol{\gamma}_{\setminus i}, \boldsymbol{\beta}, \boldsymbol{y})} = \frac{f(\boldsymbol{y} \mid \boldsymbol{\beta}, \gamma_i = 1, \boldsymbol{\gamma}_{\setminus i})}{f(\boldsymbol{y} \mid \boldsymbol{\beta}, \gamma_i = 0, \boldsymbol{\gamma}_{\setminus i})} \frac{f(\boldsymbol{\beta} \mid \gamma_i = 1, \boldsymbol{\gamma}_{\setminus i})}{f(\boldsymbol{\beta} \mid \gamma_i = 0, \boldsymbol{\gamma}_{\setminus i})} \frac{f(\gamma_i = 1, \boldsymbol{\gamma}_{\setminus i})}{f(\gamma_i = 0, \boldsymbol{\gamma}_{\setminus i})} \tag{11}$$

Note that (9) seems less natural than (4) as $f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y})$ may depend on $\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}}$. One way of avoiding this is to assume prior conditional independence of $\boldsymbol{\beta}_i$ terms given $\boldsymbol{\gamma}$, in which case $f(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ can be omitted from (9). This is a restrictive assumption but may be realistic when priors are intended to be non-informative, particularly if the columns of different $\boldsymbol{X}_i$ in (8) are orthogonal to each other. Then, each prior for $\boldsymbol{\beta}_i \mid \boldsymbol{\gamma}$ consists of a mixture of two densities. The first, $f(\boldsymbol{\beta}_i \mid \gamma_i = 1, \boldsymbol{\gamma}_{\setminus i})$, is the true prior for the parameter whereas the second, $f(\boldsymbol{\beta}_i \mid \gamma_i = 0, \boldsymbol{\gamma}_{\setminus i})$, is a pseudoprior.

Expression (11) is similar to equivalent expressions in other proposed variable selection methods. George and McCulloch (1993) propose a 'Stochastic Search Variable Selection' strategy which assumes the maximal model throughout, but constrains $\boldsymbol{\beta}_i$ parameters to be close to zero when $\gamma_i = 0$. In this situation, $f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma})$ is independent of $\boldsymbol{\gamma}$ and so the first ratio on the right hand side of (11) can be omitted. Kuo and Mallick (1998) propose a similar approach to the above but use a prior distribution for $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with $\boldsymbol{\beta}$ independent of $\boldsymbol{\gamma}$. Then, the second term on the right hand side of (11) can be omitted.

Carlin and Chib's method involves a single model indicator parameter. Therefore at each iteration of the Gibbs sampler, all parameters of all models are generated from either posterior distribution or pseudoprior, and the model selection step allows a simultaneous change of all $\gamma_i s$. For Gibbs variable selection, an observation of $\boldsymbol{\gamma}$ is generated following generation of all $\boldsymbol{\beta}_k$ from posterior distributions or pseudopriors. This procedure will generally involve generating from p conditional distributions for $\boldsymbol{\beta}_k$, a much smaller burden than required for Carlin and Chib's method. Furthermore, it would seem to be more efficient to generate pairs of $(\boldsymbol{\beta}_k, \gamma_k)$ successively, possibly by a random scan, so that more local moves in model space are attempted.

Clearly, moves between models $m(\boldsymbol{\gamma})$ and $m'(\boldsymbol{\gamma}')$ may also be based on a Metropolis step, as was suggested in Section 2.4. Then the pseudopriors may be thought of as part of the proposal density for parameters which are present in one model but not in the other. This highlights a drawback with the variable selection approaches discussed in this section, namely that parameters which are 'common' to both models remain unchanged, and therefore the procedure will not be efficient unless posterior distributions for such parameters are similar under both models.

4. Choice of proposal distributions

For all of the MCMC methods we have described above, efficient performance depends on good choice of proposal

(or pseudoprior) distributions. The importance of this depends on the extent to which the moves between models can be considered to be 'local'. A local move may be characterised as one where we might hope to make use of the current parameter values to propose plausible values for the parameters of the proposed model. Typically, a local proposal will involve generating from a distribution of dimension lower than the resulting proposed parameter vector. One area where it may be possible to utilise local moves is a variable selection approach where the proposed value of γ differs from the current value in a single component. A term is either added to or deleted from the current model. In this situation, one option is to retain the parameter values for those terms which are present in both the current and proposed models. Gibbs variable selection outlined in Section 3.1 adopts this approach as do Dellaportas and Forster (1999) in their application of reversible jump to loglinear models. The proposal (pseudoprior) distribution for an additional model parameter may be normal with mean and variance estimated using a pilot chain for the saturated model, where all parameters are present. Such an approach is more likely to be successful where the predictors are orthogonal to one another, and model parameters have a similar interpretation across model. In situations where this is not the case, then the reversible jump approach may be adapted to deal with this. For example, when a term is removed from a model, the presence of the function g allows the parameters of the proposed model to depend deterministically on those of the current model in a flexible way. See Section 6 and the example in Section 2.4 for details.

An alternative is to propose more global moves, where the relationship between the parameters of the models is less obvious, and it is more difficult to generate parameters for the proposed model which depend on the current parameter values in any sensible way. In these situations the kind of proposals utilised by the independence sampler and the equivalent Metropolised Carlin and Chib approach, may allow for more 'global' moves, which traverse B more quickly. However, this is heavily dependent on being able to generate sensible values for all the parameters of any model proposed. Carlin and Chib (1995) propose a pilot chain for each model to identify a mean and variance for a normal proposal. This is expensive when M is large.

One possible strategy, where appropriate, is to generate parameter values using a normal distribution centred at the maximum likelihood estimate for $\boldsymbol{\beta}_m$ with variance equal to the asymptotic variance of the mle. This potentially requires an expensive maximisation at for each proposed model. An alternative, for a generalised linear model with link function h, variance function v and scale parameter ϕ is to approximate the distribution of z, the vector with components $h(y_i), i = 1, \ldots, n$, by a normal distribution. Then, the posterior distribution of $\boldsymbol{\beta}_m$ may be approximated by a normal distribution with mean $(\boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T \boldsymbol{W} z$ and variance $(\boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{X}_m)^{-1}$ where $\boldsymbol{X}_m$ is the design or data matrix for model m, and $\boldsymbol{W}$ is a diagonal matrix with components $\{v(\hat{\mu}_i)h'(\hat{\mu}_i)^2\hat{\phi}_i\}^{-1}, i = 1, \ldots, n$. Here $\hat{\mu}_i$ may be the saturated model estimate $(= y_i)$ or any other appropriate estimate. Although this approximation is quite crude, it may still suffice as a proposal distribution. No maximisation is required, although matrix inversion may still prove expensive.

The general reversible jump procedure allows any level of compromise between the local and global extremes discussed above. See, for example, Richardson and Green (1997), and accompanying discussion.

5. Illustrated examples

5.1. *A logistic regression example*

To illustrate the different MCMC model choice methods, we consider a dataset analysed by Healy (1988). The data, presented in Table 1, reflect the relationship between survival, condition of the patient (more or less severe) and the treatment received (antitoxin medication or not). Suppose that we wish to compare the 5 possible logistic regression models with response variable the number of survivals and explanatory factors the patient condition and the received treatment. The full model is given by

$$Y_{jl} \sim Bin(n_{jl}, p_{jl}),$$

$$\log\left(\frac{p_{jl}}{1-p_{jl}}\right) = \mu + a_j + b_l + (ab)_{jl}, \quad j, l = 1, 2$$

where Y_{jl}, n_{jl} and p_{jl} are the number of survivals, the total number of patients and the probability of survival under level j of severity and treatment l; μ, a_j, b_l and $(ab)_{jl}$ are the model parameters corresponding to the constant term, level j of severity, treatment l, and interaction of severity j and treatment l.

We consider the Gibbs variable selection introduced in Section 3.1, Kuo and Mallick's method, local reversible jump as implemented by Dellaportas and Forster (1999), and the Metropolised version of Carlin and Chib's method (the independence sampler for model jumps) presented in Section 2.4. For the latter, we used two different forms of proposal density, the multivariate normal density suggested in Section 4 with $\hat{\mu}_i$ in the variance replaced by the observed proportions, and $\phi_i = 1/n_i$; and a proposal where the model parameters have independent $N(\bar{\beta}_i, \bar{\sigma}_i^2)$ densities, where $\bar{\beta}_i$ and $\bar{\sigma}_i^2$ were estimated from a pilot run of 500 iterations in the full model, after discarding the first 100 as burn-in iterations. The resulting values of $\bar{\beta}_i$ and $\bar{\sigma}_i$ were $(-0.47, -0.87, 0.56, -0.17)$ and $(0.27, 0.27, 0.28, 0.27)$ for $i = 0, \ldots, 3$ respectively. The same normal distributions for β_i are used as proposals the local reversible jump and as pseudopriors for Gibbs variable selection.

A rough guideline for our comparisons is the approximation to the Bayes factor $B_{m_1 m_0}$ of model m_1 against model m_0 based

Table 1. *Logistic regression example dataset*

	Antitoxin	Death	Survivals
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

on BIC,

$$-2\log B_{m_1 m_0} \simeq -2\log\left(\frac{f(\boldsymbol{y}\mid\hat{\boldsymbol{\beta}}_{m_1}, m_1)}{f(\boldsymbol{y}\mid\hat{\boldsymbol{\beta}}_{m_0}, m_0)}\right) + (d(\boldsymbol{\beta}_{m_1}) - d(\boldsymbol{\beta}_{m_0}))\log n \qquad (12)$$

where $\hat{\boldsymbol{\beta}}_m$ denotes the maximum likelihood estimator under model m and $n = \sum_{ij} n_{ij}$ for the logistic regression example; see Raftery (1996) for details.

Assuming the usual sum-to-zero constraints, the parameter vector for the full model is given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (\mu, a_2, b_2, (ab)_{22})$. We use the same $N(0, \sigma^2_{\beta_i})$ prior distribution, where relevant, for each β_i, $i = 0, \ldots, 3$, under all five models. Following the ideas of Dellaportas and Forster (1999) we choose $\sigma^2_{\beta_i} = 8$ as a prior variance which gives a diffuse but proper prior.

All the Markov chains were initiated at the full model with starting points $\beta_i = 0$ for all $i = 0, \ldots, 3$. For the reversible jump and Metropolised Carlin and Chib methods we propose a 'neighbouring' model which differs from the current model by one term, so $j(m, m) = 0$. Within each model, updating of the parameters β_i was performed via Gibbs sampling steps as described in Dellaportas and Smith (1993). Finally, each Markov chain ran for 21,000 iterations and the output summaries are based on ergodic averages taken over the final 20,000 iterations. All of the MCMC approaches took a similar length of time (around 40 seconds on a PC Pentium 100), and gave similar results, with the combined probability of the two most probable models at least 0.93. The full results are given in Table 2. Figures 1 and 2 show the evolution of the ergodic probability for the model with the highest posterior probabilities. The MCMC batch standard error when the generated samples were divided in thirty batches is also displayed for in Table 2, for the two most probable models. The differences between the posterior model probabilities calculated by the five methods can comfortably be attributed to this Monte Carlo error. For this example, Gibbs variable selection performs well, particularly given its ease of implementation.

5.2. *Simulated regression examples*

To evaluate the performance of the methods, we use a series of simulated linear regression examples, as presented by Raftery, Madigan and Hoeting (1997). The regression model can be written as $\boldsymbol{Y} \sim N_n(\boldsymbol{\eta}, \sigma^2\boldsymbol{I})$ with $\boldsymbol{\eta}$ given by (8).

For all examples we use independent N(0, 100) priors for the regression coefficients and *gamma*$(10^{-4}, 10^{-4})$ for σ^{-2}. The data generation details are given in Table 3. For all variable selection procedures we also included the constant term (noted as X_0) as a possible regressor.

We consider the same MCMC methods as in example one; Gibbs variable selection, Kuo and Mallick's method, the local

Table 2. *Posterior model probabilities for logistic regression example with approximate standard errors for the most probable models*

Model	Deviance	AP	GVS	KM	RJ	MCC(I)	MCC(M)
μ	18.656	0.004	0.004	0.008	0.005	0.004	0.005
$\mu + a_j$	4.748	0.460	0.492	0.484	0.489	0.490	0.494
Standard error			0.015	0.045	0.033	0.033	0.026
$\mu + b_l$	12.171	0.011	0.010	0.009	0.011	0.011	0.012
$\mu + a_j + b_l$	0.368	0.462	0.440	0.450	0.442	0.441	0.435
Standard error			0.014	0.041	0.027	0.027	0.020
$\mu + a_j + b_l + (ab)_{jl}$	0.000	0.063	0.053	0.050	0.053	0.054	0.054

AP: Approximate probabilities based on BIC, GVS: Gibbs variable selection, KM: Kuo and Mallick's Gibbs sampler, RJ: Reversible jump as implemented by Dellaportas and Forster (1999), MCC(I): Metropolised Carlin and Chib with Independent Pilot Run Proposals, MCC(M): Metropolised Carlin and Chib with Multivariate Proposals.

Table 3. *Simulated datasets details (n is the sample size, p is the number of variables considered excluding the constant term, design structure 1: $X_0 = 1$, $X_i \sim N(0, 1)$, $i = 1, \ldots, p$ and design structure 2: $X_0 = 1$, $X_i \sim N(0, 1)$, $i = 1, \ldots, 10$ and $X_i \sim N(0.3\,X_1 + 0.5\,X_2 + 0.7\,X_3 + 0.9\,X_4 + 1.1\,X_5, 1)$, $i = 11, \ldots, 15$*

Dataset	n	p	Design structure	Generated $\boldsymbol{\eta}$	model σ	Backward/forward selected model	MCMC best model
1	50	15	1	$X_4 + X_5$	2.50	$X_4 + X_5 + X_{12}$	$X_4 + X_5$
2	50	15	2	$\sum_{i=1}^{5} X_i$	2.50	$\sum_{i=1}^{5} X_i + X_{12}^a$	X_{14}
3	50	15	1	0	2.50	*Empty*	*Empty*
4	50	15	2	0	2.50	$X_3 + X_{12}$	X_3
5	100	50	1	0	1.00	X_{19}	*Empty*
6	100	30	1	$0.5X_1$	0.87	X_1	X_1

[a] This model was selected only by backward procedure. Forward procedure selected model X_{14}.

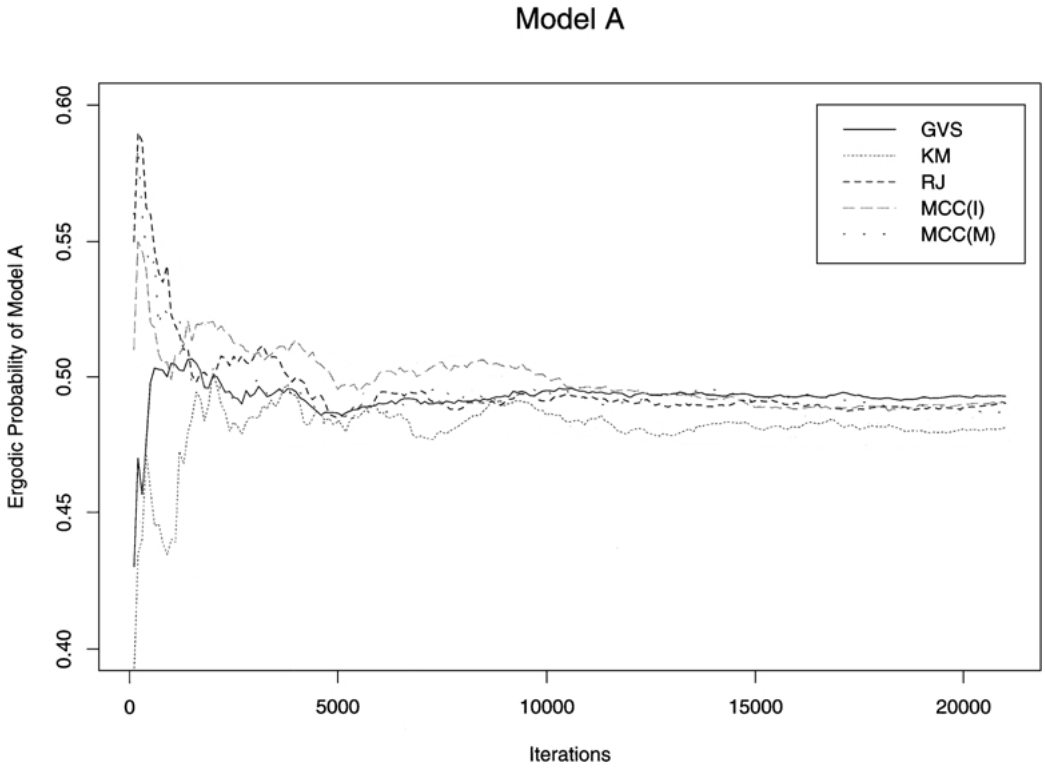

Fig. 1. *Ergodic posterior model probability of model A for logistic regression example (GVS: Gibbs variable selection, KM: Kuo and Mallick's Gibbs sampler, RJ: Reversible jump as implemented by Dellaportas and Forster (1999), MCC(I): Metropolised Carlin and Chib with independent proposals, MCC(M): Metropolised Carlin and Chib with multivariate proposals)*

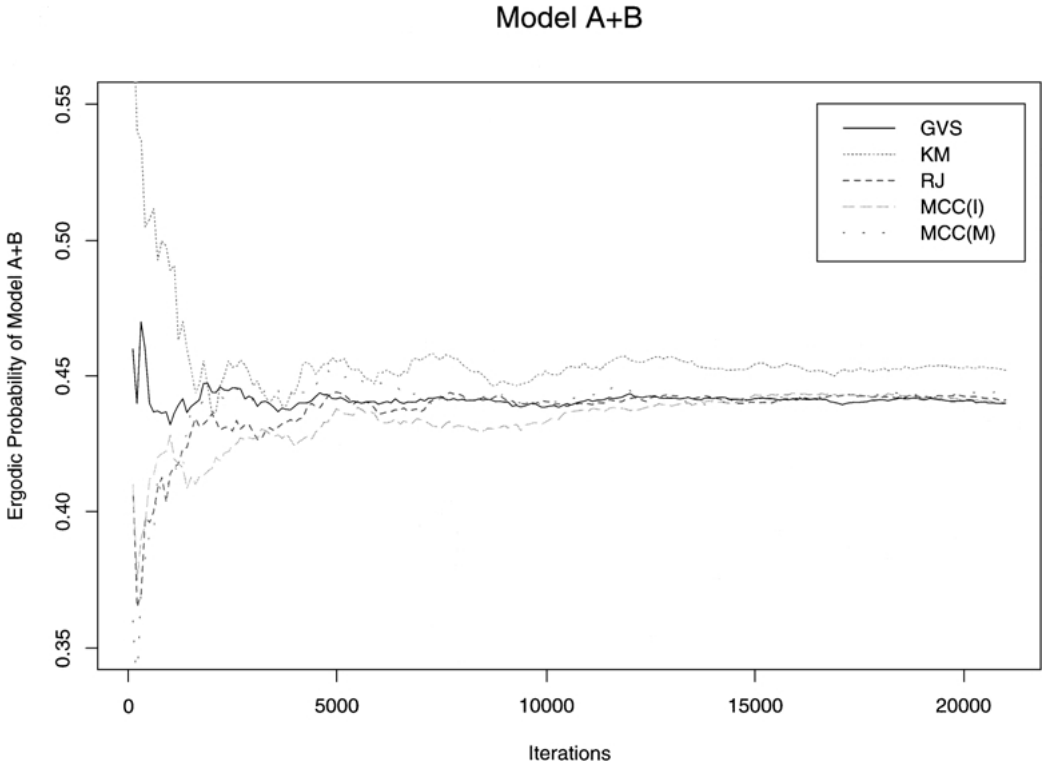

Fig. 2. *Ergodic posterior model probability of model $A + B$ for logistic regression example (GVS: Gibbs variable selection, KM: Kuo and Mallick's Gibbs sampler, RJ: Reversible jump as implemented by Dellaportas and Forster (1999), MCC(I): Metropolised Carlin and Chib with independent proposals, MCC(M): Metropolised Carlin and Chib with multivariate proposals)*

Table 4. *Batch standard deviations of highest posterior model probability for simulated regression datasets (GVS: Gibbs variable selection, KM: Kuo and Mallick's Gibbs sampler, RJ: Reversible jump as implemented by Dellaportas and Forster (1999), MCC(I): Metropolised Carlin and Chib with independent proposals, MCC(M): Metropolised Carlin and Chib with multivariate proposals)*

	Dataset					
	1	2	3	4	5	6
GVS	0.017	0.077	0.024	0.016	0.041	0.027
KM	0.039	0.059	0.032	0.037	0.089	0.059
RJ	0.042	0.102	0.032	0.028	0.062	0.062
MCC(I)	0.044	–	0.043	0.026	0.143	0.078
MCC(M)	0.048	0.042	0.044	0.029	0.103	0.090

reversible jump as implemented by Dellaportas and Forster (1999), and the Metropolised Carlin and Chib (independence sampler) using both independent and multivariate proposal densities. The proposal distributions needed for the implementation of Gibbs variable selection, reversible jump and the Metropolised Carlin and Chib method were constructed from the sample mean and standard deviation of an initial Gibbs sample of size 500 for the full model, with initial values of zero. The multivariate proposal for Metropolised Carlin and Chib was of the form

$$f(\boldsymbol{\beta}_m \mid m \neq m') \sim N\left(\left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1} \boldsymbol{X}_m^T y, \left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1} \hat{\sigma}^2\right) \qquad (13)$$

where $\hat{\sigma}^2$ is the current value for the residual variance. Note that this is the proposal suggested in Section 4, applied for normal regression models.

To compare the performance of all methods we divided the sample output taken at fixed time intervals (5, 15 and 10 minutes for datasets 1–4, 5 and 6 respectively) into 30 equal batches and reported in Table 4 the batch standard error of the highest posterior model probability. The evolution of the corresponding ergodic posterior probabilities is displayed in Fig. 3. Again, the differences between the posterior model probabilities calculated by the five methods is well within Monte Carlo error.

As would be expected, all methods gave similar results after a reasonably long run. Generally, Gibbs variable selection seems to have lower batch standard error which indicates greater efficiency. Metropolised Carlin and Chib, although potentially more efficient per iteration loses out in comparison, due to the time taken at each iteration to propose a complete set of new parameter values. Kuo and Mallick's method generally performs worse than Gibbs variable selection but reasonably well in general.

For dataset 2, where predictors are correlated, the Metropolised Carlin and Chib sampler with the multivariate proposal distribution performed better than the other methods. This is not surprising, as coefficient values may change dramatically across models. The Metropolised Carlin and Chib sampler with independent proposals, based on the full model, also performed poorly here. It did not visit the model selected by the other methods. However, after ten million iterations this method did eventually support the same model.

6. Discussion

Efficient MCMC investigation of posterior distributions is highly dependent on good choice of transition distribution. This seems particularly true in the presence of model uncertainty. There is a natural trade-off between local transitions and more global transitions which potentially allow one to explore the distribution of interest more quickly, but may be much harder to implement efficiently.

In this paper, we have focussed on a number of MCMC approaches which have been proposed for investigating model uncertainty. These range from highly local approaches which restrict transitions to 'neighbouring' models, to completely global approaches where any transition is, in principle, possible. We believe that simple local approaches can be successful and the illustrative examples presented in this paper bear this out. The local version of reversible jump (as implemented by Dellaportas and Forster 1999) and Gibbs variable selection are effective when parameters have a similar interpretation across models, and hence marginal posterior densities do not change greatly.

On the other hand, the Metropolised version of Carlin and Chib's Gibbs sampler, equivalent to the independence sampler for model jumps, is more effective in cases where interpretation of model parameter changes between models. Ideally, this involves the construction of proposal or pseudoprior densities that are good approximations of the target posterior distribution for each model. This is an extra computational burden, which becomes larger as the number of possible models increases. However, the same algorithm with model parameters proposed independently of one another requires only the same amount of 'training' as Gibbs variable selection or our local version of reversible jump, and the examples illustrate that this method can also be effective.

One possible strategy for generalised linear models, is to examine the approximate posterior correlation matrix of the model parameters under the full model. If all correlations are low then we can directly use either simple local reversible jump or Gibbs variable selection, with independent proposals (pseudopriors) with parameters taken from a pilot run of the full model. If high correlations exist then a more global method such as Metropolised Carlin and Chib (independence sampler) with multivariate proposal densities may be preferred. An alternative, in the presence of high posterior correlations is to consider a more sophisticated reversible jump sampler. For example, when moving to a nested model m' of lower dimension, a possible proposal is

$$\boldsymbol{\beta}_{m'} = \left(\boldsymbol{X}_m^T \boldsymbol{W} \boldsymbol{X}_{m'}\right)^{-1} \boldsymbol{X}_{m'}^T \boldsymbol{W} \boldsymbol{X}_m \boldsymbol{\beta}_m.$$

Then, when considering a transition from m' to m, the proposed $\boldsymbol{\beta}_m$ must be one for which the above equation holds. This approach is local in the sense that it involves a minimal

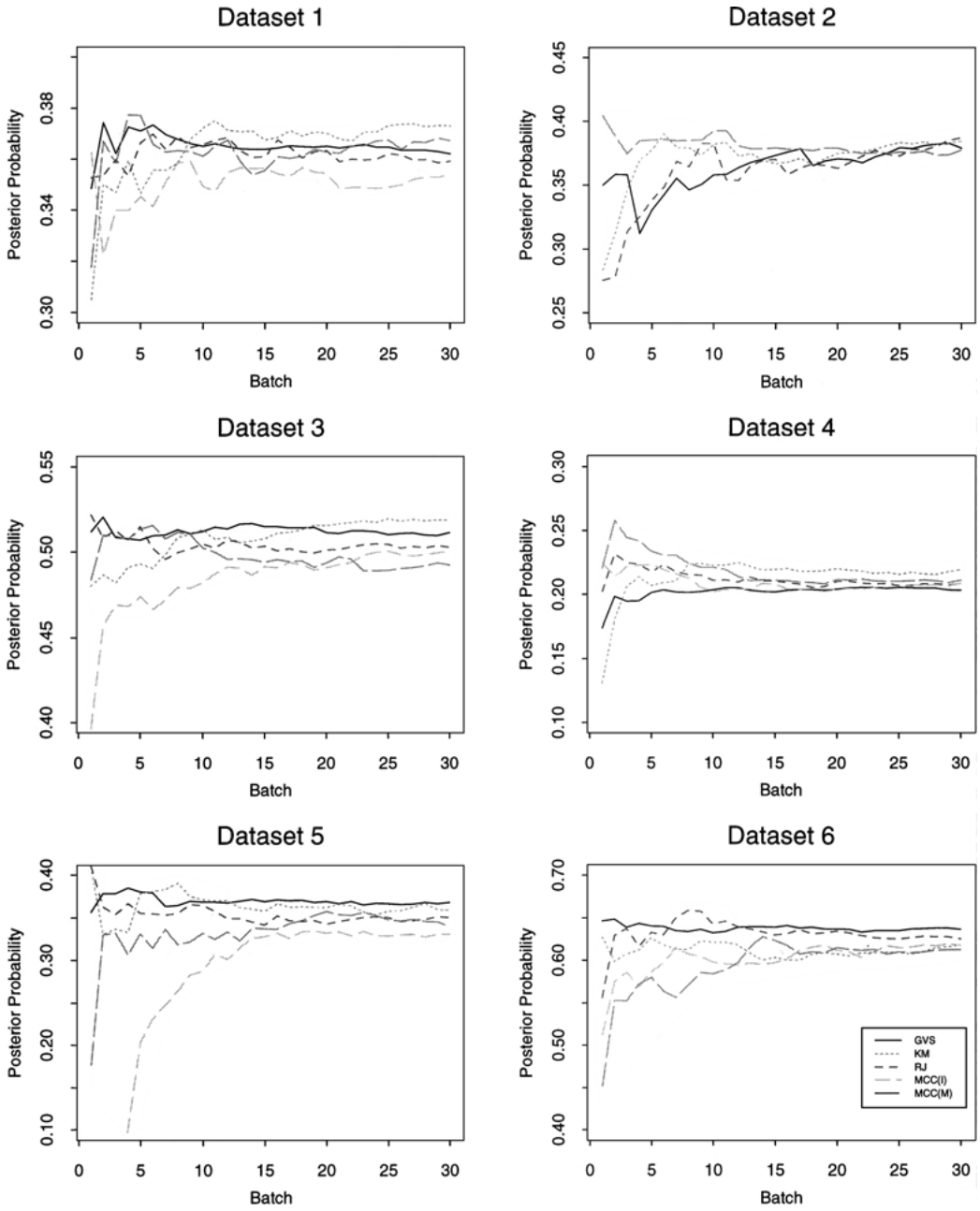

Fig. 3. *Ergodic highest posterior model probabilities for simulated datasets (GVS: Gibbs variable selection, KM: Kuo and Mallick's Gibbs sampler, RJ: Reversible jump as implemented by Dellaportas and Forster (1999), MCC(I): Metropolised Carlin and Chib with independent proposals, MCC(M): Metropolised Carlin and Chib with multivariate proposals)*

change to the linear predictor (an object whose interpretation is model-independent) even if the parameters for terms which are present in both models may change greatly.

Acknowledgments

We would like to acknowledge helpful comments from Brad Carlin, Sid Chib, two anonymous referees, and participants at a workshop on 'Variable dimension MCMC for Bayesian model choice applications' supported by the European Science Foundation Highly Structured Stochastic Systems Network.

References

Besag J. 1997. Comment on 'On Bayesian analysis of mixtures with an unknown number of components'. Journal of Royal Statistical Society B 59: p. 774.

Carlin B.P. and Chib S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of Royal Statistical Society B 157: 473–484.

Chipman H. 1996. Bayesian variable selection with related predictors. Canadian Journal of Statistics 24: 17–36.

Clyde M. and DeSimone-Sasinowska H. 1998. Accounting for model uncertainty in Poisson regression models: Does particulate matter particularly matter? Institute of Statistics and Decision Sciences, Duke University, Technical Report 97–06.

Dellaportas P. and Forster J.J. 1999. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86: 615–633.

Dellaportas P. and Smith A.F.M. 1993. Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. Applied Statistics 42: 443–459.

George E. and McCulloch R.E. 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881–889.

Godsill S.J. 1998. On the relationship between MCMC model uncertainty methods. Signal Processing Group, Cambridge University Engineering Department, Technical Report.

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

Green P. and O'Hagan A. 1998. Model choice with MCMC on product spaces without using pseudopriors. Department of Mathematics, University of Nottingham, Technical Report.

Gruet M.-A. and Robert C. 1997. Comment on 'On Bayesian analysis of mixtures with an unknown number of components'. Journal of Royal Statistical Society B 59: p. 777.

Healy M.J.R. 1988. Glim: An Introduction. Claredon Press, UK.

Kuo L. and Mallick B. 1998. Variable selection for regression models. Sankhyā, B 60: 65–81.

Madigan D. and York J. 1995. Bayesian graphical models for discrete data. International Statistical Review 63: 215–232.

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

Raftery A.E., Madigan D., and Hoeting J.A. 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92: 179–191.

Richardson S. and Green P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of Royal Statistical Society B 59: 731–792.

Roberts G.O. 1996. Markov chain concepts related to sampling algorithms. In: Gilks W.R., Richardson S., and Spiegelhalter D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London, pp. 45–57.

Tierney L. 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22: 1701–1762.