

ON THE RELATIONSHIP BETWEEN MCMC
MODEL UNCERTAINTY METHODS

BY SIMON J. GODSILL

Signal Processing Group,
Cambridge University Engineering Department,
Trumpington Street
Cambridge CB2 1PZ, UK
email: sgj@eng.cam.ac.uk

Also available as:
Tech. Report CUED/F-INFENG/TR.305, Nov.
1997

On the relationship between MCMC model uncertainty methods

BY SIMON J. GODSILL

*Signal Processing Group,
Cambridge University Engineering Department,
Cambridge CB2 1PZ, UK*

sjg@eng.cam.ac.uk

September 17, 1998

SUMMARY

We discuss relationships between the existing methods for MCMC exploration of model spaces, including the reversible jump sampler of Green (1995), the ‘model composition’ approach of Carlin and Chib (1995), the MC³ techniques of Madigan and Raftery (1995) and MCMC methods for variable selection such as George and McCulloch (1993), Kuo and Mallick (1997) and Geweke (1996). We link these different methods together through a composite model space similar to that used by Carlin and Chib in which a model of constant dimensionality is created by considering the product space of parameters from all possible models within the candidate set and the model indexing variable. In the examples given in their paper, Carlin and Chib apply a straightforward Gibbs sampler to the composite space which renders the method impracticable for comparison between more than a small handful of models. We show that the other methods of MCMC model selection can be obtained by applying different forms of MCMC sampling to the composite space. The results shed some light upon the issues of ‘pseudo-prior’ selection in the case of the Carlin and Chib sampler and choice of proposal distribution in the case of Green’s reversible jump method. Furthermore, we propose efficient reversible jump proposal schemes which take advantage of any analytic structure that may be present in the model. The method is compared

with a standard reversible jump scheme for the problem of model order uncertainty in autoregressive time series.

Some key words: Model selection; MCMC; Reversible jump; Variable selection

1 Introduction

Within a Bayesian setting model uncertainty can be handled in a parametric fashion through the use of posterior model probabilities. Suppose there exist M candidate models, one of which is assumed to be a perfect statistical description of an observed data vector y . Associated with each model is a likelihood $p(y|\psi_k, k)$ that depends upon an (unknown) set of parameters ψ_k , where $k \in \{1, \dots, M\}$ denotes the k th model in the list of candidates. In general ψ_k may be multivariate and may have different dimensionality and support Ψ_k in different models. A prior distribution $p(\psi_k|k)$ is assigned to each parameter vector and a prior distribution $p(k)$ to the model number, reflecting prior knowledge about the probabilities of individual models. The posterior model probability for model k is then obtained as:

$$p(k|y) = \frac{p(y|k)p(k)}{p(y)} = \frac{\int_{\Psi_k} p(y|\psi_k, k)p(\psi_k|k)d\psi_k p(k)}{p(y)}.$$

The term $p(y|k)$ is sometimes referred to as the *marginal likelihood* for model k . We assume throughout that the parameter priors $p(\psi_k|k)$ are proper.

In some cases the goal of statistical analysis may simply be to summarise the relative posterior probabilities of the individual models or to estimate a single ‘best’ model through the use of some suitable risk function. In many applied scenarios, however, model uncertainty will be incorporated into tasks such as forecasting, interpolation, smoothing or noise reduction (West & Harrison, 1997). If we denote the required quantity as ϕ , where ϕ may be the data points for forecast/interpolation or simply some parameter of interest which is common to all models, then we can obtain the posterior probability for ϕ by ‘model averaging’:

$$p(\phi|y) = \sum_k p(\phi, k|y) = \frac{\sum_k \int_{\Psi_k} p(y|\phi, \psi_k, k)p(\phi|\psi_k, k)p(\psi_k|k)d\psi_k p(k)}{p(y)}. \quad (1)$$

Calculation of posterior model probabilities is rarely achievable in analytic form for realistic models. Approximation methods may be used, and there is a large array of

tools available (see e.g. Raftery (1996) for a good review coverage). One effective means of achieving this is through a Monte Carlo sampling scheme. If one can draw random samples (ϕ^i, ψ_k^i, k^i) from the joint posterior distribution $p(\phi, \psi_k, k|y)$ then Monte Carlo estimates can be made for any unknown posterior quantities, for example:

$$E[g(\phi)|y] \approx \frac{1}{K} \sum_{i=1}^K g(\phi^i)$$

where $g()$ is some arbitrary functional whose posterior expectation is required.

For distributions of parameters with fixed dimensionality a suitable scheme for drawing a dependent sequence of samples from the joint posterior is Markov chain Monte Carlo (MCMC). MCMC methods (Metropolis et al., 1953, Geman & Geman, 1984, Gelfand & Smith, 1990, Hastings, 1970) have become well established over recent years as a powerful tool for analysis of complex statistical problems. Until relatively recently, however, these methods were applied only to problems with fixed dimensionality. We consider in this paper relationships between some existing methods for dealing with Bayesian model uncertainty using MCMC methods. The MCMC model combination (MC³) methods of Madigan & Raftery (1995), developed in the context of decomposable graphical models, show how to explore model space when the marginal likelihood $p(y|k)$ is available analytically. This applies to relatively few realistic models and hence does not find very general application. As an alternative, MCMC variable selection methods provide flexible methodology for models which can be parameterised into individual explanatory variables which are switched in or out depending upon their relevance to the observed data. In particular, George & McCulloch (1993) have developed stochastic search variable selection (SSVS) methods in which the prior model structure encourages small parameter values for variables which do not explain the data well. The variable selection problem as posed by George and McCulloch is inherently of fixed dimensionality and hence standard MCMC procedures such as the Gibbs Sampler can be applied in this case. These methods, and variants in which the ‘unused’ model parameters are forced exactly to zero (Geweke, 1996; Kuo & Mallick, 1997) (a problem of varying dimensionality), have found application in a number of applied areas including outlier analysis (McCulloch & Tsay, 1994; Godsill & Rayner, 1996; Godsill, 1997; Barnett et al., 1996) and time series model selection (Barnett et al., 1996; Troughton & Godsill, 1997a; Huerta & West, 1997), to name only a few. Further relevant papers on Bayesian variable selection are by Clyde et al. (1996), Hoeting et al. (1996) and Raftery et al. (1997).

A more general technique is presented by Carlin & Chib (1995) in which model uncertainty is expressed through a probability measure over the product space of all possible model parameters. This technique has an appealing simplicity and generality. However, in the Gibbs sampler form in which it is demonstrated in the paper, the method is not easily applicable to problems with more than a handful of competing models, owing to the necessity for choosing and generating from a large number of ‘pseudo-priors’. Naturally, it is possible to apply other forms of MCMC to the same product space model, and it is this idea which we use to demonstrate the close relationship between Carlin and Chib’s framework and the other existing MCMC methods for dealing with model uncertainty.

The most flexible methods to date are the reversible jump sampler (Green, 1995) and the related jump diffusion methods (Grenander & Miller, 1991; Grenander & Miller, 1994; Phillips & Smith, 1994), which can be implemented in principle for any type of model uncertainty. These methods have been successfully applied in many areas including mixture modelling (Richardson & Green, 1997), image analysis (Green, 1995; Morris, 1996), changepoint analysis (Green, 1995) and time series model selection (Barbieri & O’Hagan, 1996; Troughton & Godsill, 1997b), again listing only a few of the many contributions in this area.

In this paper it is shown that the existing MCMC methods for model selection can be linked together through consideration of a composite model space similar to that of Carlin & Chib (1995). This sheds some light upon the issues of pseudo-prior choice, in the case of Carlin and Chib, and choice of proposal densities in the case of Green’s reversible jump sampler. It may also lead to new classes of model space sampler which combine the benefits of several different schemes.

2 Relationships between different model space samplers

2.1 A composite representation of model uncertainty problems

We define firstly a composite model space which accounts for the types of model uncertainty we need to consider. This composite model is a straightforward generalisation

of that developed by Carlin & Chib (1995). Consider a ‘pool’ of N parameters $\theta = (\theta_1, \dots, \theta_N)$, such that θ_i has support Θ_i . The parameters θ_i may once again be vectors of differing dimensionality and one or more of them may have dimension zero (this would occur, for example, in the case where simple hypotheses with no unknown parameters are to be compared or in variable selection for the hypothesis that the data do not depend on any of the candidate variables). A candidate model k in the notation of the previous section can be described in terms of this pool of parameters by means of an indexing set $\mathcal{I}(k) = \{i_1(k), i_2(k), \dots, i_{l(k)}(k)\}$ which contains $l(k)$ integer values between 1 and N . ψ_k in the earlier notation is then obtained as $\psi_k = (\theta_i; i \in \mathcal{I}(k))$ and $\Psi_k = \prod_{i \in \mathcal{I}(k)} \Theta_i$. In the simplest case we have $\mathcal{I}(k) = k$, which leads to a one-to-one correspondence between the two formulations. In other cases such as variable selection or nested models it may be convenient to ‘share’ parameters between more than one model. A probability distribution is now defined over the entire product space of candidate models and their parameters, i.e. $(k, \theta) \in \mathcal{K} \times \prod_{i=1}^N \Theta_i$, where \mathcal{K} is the set of candidate model indices. The likelihood and prior structure are then defined in a corresponding way, as follows. For a particular k the set $\mathcal{I}(k)$ defines the conditional dependence structure of model k , i.e.

$$p(y|k, \theta) = p(y|k, \theta_{\mathcal{I}(k)}) \quad (2)$$

where $\theta_{\mathcal{I}(k)} = (\theta_i; i \in \mathcal{I}(k))$ denotes the parameters used by model k . The model specification is completed by the parameter prior $p(\theta_{\mathcal{I}(k)}|k)$ and the model prior $p(k)$. The full posterior distribution for the composite model space can now be expressed as

$$p(k, \theta|y) = \frac{p(y|k, \theta_{\mathcal{I}(k)}) p(\theta_{\mathcal{I}(k)}|k) p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k) p(k)}{p(y)} \quad (3)$$

where $\theta_{-\mathcal{I}(k)} = (\theta_i; i \in \{1, \dots, N\} - \mathcal{I}(k))$ denotes the parameters *not* used by model k . All of the terms in this expression are defined explicitly by the chosen likelihood and prior structures except for $p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k)$, the ‘prior’ for the parameters in the composite model which are not used by model k . It is easily seen that any proper distribution can be assigned arbitrarily to these parameters without affecting the required marginals for the remaining parameters. We have given the general case, in which this prior can depend upon both k and the remaining model parameters. In many cases it will be convenient to assume that the unused parameters are *a priori* independent of one another and also of $\theta_{\mathcal{I}(k)}$. In this case we have that $p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k) = p(\theta_{-\mathcal{I}(k)}|k) = \prod_{i \notin \mathcal{I}(k)} p(\theta_i|k)$ and the composite model posterior

can be rewritten as:

$$p(k, \theta|y) = \frac{p(y|k, \theta_{\mathcal{I}(k)}) p(\theta_{\mathcal{I}(k)}|k) \left(\prod_{i \notin \mathcal{I}(k)} p(\theta_i|k) \right) p(k)}{p(y)}. \quad (4)$$

Carlin & Chib (1995) presented this form of the composite representation for the case where $\mathcal{I}(k) = k$. The priors on the unused parameters $\theta_{-\mathcal{I}(k)}$ are referred to as ‘pseudo-priors’ or linking densities in the Carlin and Chib model, appropriate choice of which is crucial to the effective operation of their algorithm. We will retain the general form as given in equation (3), referring to the term $p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k)$ as the pseudo-prior, although it should be noted that the simpler form of equation (4) will often be used in practice, with consequent simplification of the posterior distribution.

In general it will be possible to parameterise the standard model uncertainty scenarios in many equivalent ways using the composite model. Some convenient parameterisations are listed below:

- *Standard model selection.* In the basic model selection problem we associate one parameter vector with each model, so we can simply use $k \in \{1, \dots, N\}$ and $\mathcal{I}(k) = \{k\}$. The model is assumed not to be nested, so no elements of different parameter vectors are considered to be common across different models. Of course all model uncertainty problems can be formulated in this way, but it will often not be convenient to use this representation for computational reasons.
- *Nested models.* In nested models it is assumed that parameters from the model of order k have the same interpretation as the first k parameters in the model of order $k + 1$. In this case we have $k \in \{1, \dots, N\}$, as before, but now $\mathcal{I}(k) = \{1, \dots, k\}$.
- *Variable selection.* In variable selection problems the model conditional likelihood can depend upon any combination of the available parameters. In this case a natural parameterisation for k is as a binary N -vector, i.e. $k = [k_1, k_2, \dots, k_N] \in \{0, 1\}^N$, and $\mathcal{I}(k) = \{i : k_i = 1\}$. Each element of k then ‘switches’ a particular dependent variable in or out of the model (for example, setting $k = [0, \dots, 0]$ corresponds to the case where the data depend upon none of the candidate variables). This is a pure variable selection problem in which the model conditioned likelihood is truly independent of all those candidate variables which have $k_i = 0$. Many problems which involve latent indicator variables can be viewed as variable selection problems, and we thus use the term here in its most general sense to include all of these variants on the problem. This parameterisation of the variable selection problem is equivalent to that used by Kuo & Mallick (1997) and

Geweke (1996). Note that the stochastic search variable selection (SSVS) methods of George & McCulloch (1993) are not quite the same as this since the likelihood in their case is of fixed form for all models, depending upon all parameters within every model. Model uncertainty is then built in through a prior structure which enforces very small values for those parameters which are switched ‘off’ in the model. This avoids some of the difficulties of working with a variable dimension parameter space. Within the framework of the composite model we could easily achieve this configuration by setting $\mathcal{I}(k) = \{1, \dots, N\} \forall k$. The likelihood (2) is then taken as independent of k and distinction between different k is achieved purely through the prior distributions on the θ_i ’s and k . We will consider here the first formulation in which parameters can be switched out of the model completely, i.e. the likelihood is completely independent of θ_i when $k_i = 0$. Methods based upon these principles find application not only in traditional ‘variable selection’ problems but in many other areas where individual effects can be modelled via latent indicator variables (see references in section 2.3).

The key feature of the composite model space is that the dimension remains fixed even when the model number k changes. This means that standard MCMC procedures, under the usual convergence conditions, can be applied to the problem of model uncertainty. For example, a straightforward Gibbs sampler applied to the composite model leads to the most basic form of Carlin and Chib’s method or simple Gibbs variable selection, while we show that a more sophisticated Metropolis-Hastings approach leads to reversible jump. In the following sections we show how to obtain these existing MCMC model space samplers as special cases of the composite space sampler.

2.2 Carlin and Chib

The composite model we have described is essentially the same as that developed by Carlin & Chib (1995), although the model parameterisation in terms of $\mathcal{I}(k)$ is more flexible in that it allows parameters to be ‘shared’ between different models, which is often a desirable feature. The sampling algorithm they use to illustrate their method is easily obtained by applying a Gibbs sampler to the individual parameters θ_i and to the model index k . The sampling steps, which may be performed in a random or

deterministic scan, are as follows:

$$\theta_i \sim p(\theta_i | \theta_{-i}, k, y) \propto \begin{cases} p(y|k, \theta_{\mathcal{I}(k)}) p(\theta_{\mathcal{I}(k)}|k), & i \in \mathcal{I}(k) \\ p(\theta_{-\mathcal{I}(k)} | \theta_{\mathcal{I}(k)}, k), & i \notin \mathcal{I}(k) \end{cases}$$

$$k \sim p(k|\theta, y) \propto p(y|k, \theta_{\mathcal{I}(k)}) p(\theta_{\mathcal{I}(k)}|k) p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k) p(k).$$

The method is rather impractical for problems with many candidate models since every parameter vector is sampled at each iteration, although Green & O’Hagan (1997) have shown that this is in fact not necessary for strict convergence of the sampler. Furthermore, suitable choice of pseudo-priors is essential for efficient operation. Carlin and Chib suggest the use of pseudo-priors which are close to the posterior conditional for each model. We can see why this might be a good choice by analysing the case when the pseudo-priors are set *exactly* to the posterior conditionals for each parameter and the individual parameters are assumed independent *a priori* in the basic model selection scenario ($\mathcal{I}(k) = k$), i.e.

$$p(\theta_{-\mathcal{I}(k)}|k) = \prod_{i \notin \mathcal{I}(k)} p(\theta_i|y, k = i).$$

The sampling step for k is then found to reduce to

$$k \sim p(k|\theta, y) = p(k|y) = \int_{\Theta_k} p(\theta_k, k|y) d\theta_k.$$

In other words the model index sampling step becomes simply a draw from the true model posterior probability distribution $p(k|y)$ and does not depend upon the sampled parameter values θ_i . This is in some sense the ideal case since the aim of model uncertainty sampling is to design a sampler which explores model space according to $p(k|y)$. We can see then why choosing pseudo-priors which are close to the parameter conditionals is likely to lead to effective operation of the algorithm. Of course, the exact scheme is impractical for most models since $p(k|y)$ is typically unavailable in closed form.

2.3 MCMC Variable Selection

Using the parameterisation described above in which k is a binary vector of parameter ‘indicators’, MCMC variable selection methods are obtained immediately by the application of a Gibbs sampler to the parameter space partitioned as $(k_1, k_2, \dots, k_N, \theta_1, \theta_2, \dots, \theta_N)$. If, for simplicity, we omit any additional hyperparameters such as noise variances, which

are often considered to be common to all models, then the following sampling scheme is obtained, which is essentially the same as that of Kuo & Mallick (1997):

$$\theta_i \sim p(\theta_i | \theta_{-i}, k, y) \propto \begin{cases} p(y | \theta_{\mathcal{I}(k)}, k) p(\theta_{\mathcal{I}(k)} | k) p(\theta_{-\mathcal{I}(k)} | \theta_{\mathcal{I}(k)}, k), & k_i = 1 \\ p(\theta_i | \theta_{-i}, k), & k_i = 0 \end{cases}$$

$$k_i \sim p(k_i | k_{-i}, \theta, y) \propto p(y | \theta_{\mathcal{I}(k)}, k) p(\theta_{\mathcal{I}(k)} | k) p(\theta_{-\mathcal{I}(k)} | \theta_{\mathcal{I}(k)}, k)$$

The individual parameters θ_i are thus sampled either from their posterior conditional or from their pseudo-prior, depending upon the value of k_i . See Kuo & Mallick (1997) and Dellaportas et al. (1997) for examples of these methods applied to standard variable selection problems. Clearly schemes which use other types of MCMC in the moves or choose alternative blocking strategies to yield improved performance can also be devised (see e.g. Godsill & Rayner (1996), Barnett et al. (1996), Carter & Kohn (1996), Troughton & Godsill (1997a) and Liu (1996)).

The fact that the pseudo-priors can be chosen arbitrarily in exactly the same way as for the standard model selection problem is not often noted within a variable selection framework. One practically useful example of this fact, in some variable selection models, such as those involving linear conditionally Gaussian assumptions for the parameters, is to choose the pseudo-prior for each parameter θ_i to be the conditional posterior for θ_i with $k_i = 1$, i.e. set:

$$p(\theta_i | k_i = 0, \theta_{-i}, k_{-i}) = p(\theta_i | k_i = 1, \theta_{-i}, k_{-i}, y).$$

In the basic Gibbs sampling framework summarised above, the sampling step for k_i then reduces to:

$$k_i \sim p(k_i | \theta_i, \theta_{-i}, k_{-i}, y) = p(k_i | \theta_{-i}, k_{-i}, y) = \int_{\theta_i} p(k_i, \theta_i | \theta_{-i}, k_{-i}, y) d\theta_i.$$

When associated with the conditional draw of θ_i from its conditional posterior $p(\theta_i | \theta_{-i}, k, y)$ we see that the approach is equivalent to a blocking scheme which draws jointly for (θ_i, k_i) using the decomposition $p(\theta_i, k_i | \theta_{-i}, k_{-i}, y) = p(\theta_i | k, y) p(k_i | \theta_{-i}, k_{-i}, y)$. Such blocking schemes have been found empirically to give much improved performance over straightforward single-move Gibbs samplers both in outlier analysis (Godsill & Rayner, 1996; Godsill, 1997; Barnett et al., 1996) and variable selection for non-linear time series (Troughton & Godsill, 1997a). This blocking procedure can also be viewed as equivalent to that used by Geweke (1996), who reparameterises the problem with δ -functions in the prior for variables which are not used in the model. In these cases

the integral required can easily be performed analytically. In other cases, improved performance could be achieved over the ‘single move’ Gibbs Sampler by setting the pseudo-priors to some suitable approximation to the conditional posterior in a similar fashion to Carlin and Chib’s proposal for the basic model selection problem.

2.4 Reversible jump

In this section we consider only the basic model selection scenario in which $\mathcal{I}(k) = k$, i.e. model k uses parameters θ_k . This avoids some cumbersome notation, although it should be noted that all the methods described here are readily adapted to the more general case, including the variable selection problem.

The reversible jump sampler (Green, 1995) achieves model space moves by Metropolis-Hastings proposals with an acceptance probability that is designed to preserve detailed balance within each move type. Suppose that we propose a move to model k' with parameters $\theta'_{k'}$ from model k with parameters θ_k using a proposal distribution $q(k', \theta'_{k'}; k, \theta_k)$. The acceptance probability in order to preserve detailed balance is given by:

$$\alpha = \min \left(1, \frac{p(k', \theta'_{k'} | y) q(k, \theta_k; k', \theta'_{k'})}{p(k, \theta_k | y) q(k', \theta'_{k'}; k, \theta_k)} \right).$$

This form of acceptance probability is slightly different from the illustrations of the method given by Green (1995) in that it explicitly includes the probability of proposing the move from k to k' and that the proposal is made directly in the new parameter space $\theta'_{k'}$, rather than via ‘dimension matching’ random variables u and u' . In fact, what we are really doing is specifying the dimension-matching transformations as $\theta'_{k'} = u$ and $\theta_k = u'$, which has unity Jacobian. We thus avoid the need for a Jacobian term in the acceptance probability. This direct formulation is simply another form of Green’s method since the Jacobian in Green’s formulation arises purely as a result of the change of variables from (u', θ_k) to $(\theta'_{k'}, u)$. Note that different types of model moves from k to k' can easily be incorporated into the proposal $q()$ as components of a mixture distribution. The jump diffusion methods for model uncertainty (Grenander & Miller, 1991; Grenander & Miller, 1994; Phillips & Smith, 1994) can be considered as a special version of the reversible jump scheme, so we do not address these further here.

2.4.1 Reversible Jump derived from the composite model.

We now show that Green's reversible jump sampler can be obtained by applying a Metropolis-Hastings (M-H) proposal to the composite model space. Note that some related ideas have recently been presented independently by Besag (1997) and Dellaportas et al. (1997)¹. For simplicity we restrict ourselves to the basic model uncertainty scenario with $\mathcal{I}(k) = k$, although the ideas apply equally to variable selection and other model uncertainty problems. Consider a proposal from the current state of the composite model (k, θ) to a new state (k', θ') that takes the form:

$$q(k', \theta'; k, \theta) = q_1(k'; k) q_2(\theta'; \theta_k) p(\theta'_{-k'} | \theta'_{k'}, k').$$

This proposal, which forms a joint distribution over all elements of k' and θ' , is split into three component parts: the model index component $q_1(k'; k)$, which proposes a move to a new model index, k' ; a proposal for the parameters used by model k' , $q_2(\theta'; \theta_k)$, and a proposal on the remaining unused parameters which is equal to the pseudo-prior in model k' , $p(\theta'_{-k'} | \theta'_{k'}, k')$. We thus have a joint proposal across the whole state space of parameters and model index which satisfies the Markov requirement of the M-H method as it depends only upon the current state (k, θ) to make the joint proposal (k', θ') , and of course there are no worries about a parameter space with variable dimension since the composite model retains constant dimensionality whatever the value of k .

The acceptance probability for this special form of proposal is given, using the standard M-H procedure, by:

$$\begin{aligned} \alpha &= \min \left(1, \frac{q(k, \theta; k', \theta') p(k', \theta' | y)}{q(k', \theta'; k, \theta) p(k, \theta | y)} \right) \\ &= \min \left(1, \frac{q_1(k; k') q_2(\theta_k; \theta'_{k'}) p(\theta_{-k} | \theta_k, k) p(k', \theta'_{k'} | y) p(\theta'_{-k'} | \theta'_{k'}, k')}{q_1(k'; k) q_2(\theta'_{k'}; \theta_k) p(\theta'_{-k'} | \theta'_{k'}, k') p(k, \theta_k | y) p(\theta_{-k} | \theta_k, k)} \right) \\ &= \min \left(1, \frac{q_1(k; k') q_2(\theta_k; \theta'_{k'}) p(k', \theta'_{k'} | y)}{q_1(k'; k) q_2(\theta'_{k'}; \theta_k) p(k, \theta_k | y)} \right). \end{aligned}$$

This last line is exactly the acceptance probability for the reversible jump sampler with the proposal distribution factored in an obvious way into two components $q_1(\cdot)$ and $q_2(\cdot)$. We see that the result is independent of the value of any parameters which are unused by both models k and k' (their pseudo-priors cancel in the acceptance probability) and hence the sampling of these is a 'conceptual' step only which need not

¹The latter work, like ours, was presented for the first time at the HSSS Workshop on Model Uncertainty, New Forest, Sept. 1997

be performed in practice. This feature is a great strong-point of the method compared with the Gibbs sampling version of the Carlin and Chib method, which requires samples for all parameters in all models at every iteration. It is interesting, however, to see that both schemes can be derived as special cases of the composite space sampler.

2.4.2 Proposing from full posterior conditionals and MC³

In a similar vein to the suggestions made above for the Carlin and Chib method, a possible version of reversible jump would use the full posterior conditional $p(\theta_{k'}|k', y)$ as the proposal density $q_2(\cdot)$ in the above description. We can then employ the identity $\frac{p(k, \theta|y)}{p(\theta|k, y)} = p(k|y)$ (for example, Besag (1989) has used this basic identity to find prediction densities, Chib (1995) and Chib & Greenberg (1998) use a related identity to calculate Bayes factors) to obtain the following acceptance probability:

$$\alpha = \min \left(1, \frac{p(k'|y)q_1(k; k')}{p(k|y)q_1(k'; k)} \right).$$

This can be recognised as the acceptance probability of a standard Metropolis Hastings method with the posterior model probability $p(k|y)$ as the target distribution and using proposals $q_1(k'|k)$ for the model moves (Dellaportas et al. (1997) have independently noted this point). Note that the acceptance probability is independent of parameter values, depending only upon the proposal distribution for model order and the posterior odds ratio $p(k'|y)/p(k|y)$. Inference about parameter values can then be made conditional upon the current model index k using standard MCMC. Such a scheme has been used for decomposable graphical models in the MC³ method of Madigan & Raftery (1995). Stark et al. (1997) have also suggested a similar scheme for use with changepoint models. They point out that the parameters generated in proposing from the full conditional distribution $p(\theta_{k'}|k', y)$ can be used in a subsequent Gibbs sampling step for $\theta_{k'}$ if the move to model k' is accepted.

In the (relatively rare) cases where $p(k|y)$ or equivalently the value of the full conditional $p(\theta_k|k, y)$ at all values of θ_k is available analytically², use of conditional parameter distributions as reversible jump proposals would lead to excellent exploration of model space. This would suggest that reversible jump proposals should be designed to approximate as closely as possible the parameter conditionals in order to come close to the performance of the scheme when parameter conditionals are readily available in

²since full knowledge of $p(\theta_k|k, y)$, including its normalising constant, generally implies knowledge of $p(k|y)$

exact form. This is similar in principle to Carlin and Chib’s suggestion that pseudo-priors be chosen close to the parameter conditionals in their method.

2.4.3 Use of partial analytic structure in reversible jump proposals

The exact scheme of the last section is, of course, not available for most models of practical interest. Nevertheless, many useful models will have what we term *partial analytic* structure, that is we have the full conditional in closed form for some sub-vector of $\theta_{k'}$, the vector of parameters which are used by the new model k' ; in other words $p((\theta_{k'})_{\mathcal{U}} | (\theta_{k'})_{-\mathcal{U}}, k', y)$ is available for some subset of the parameters, indexed by a set \mathcal{U} . If we suppose that an equivalent subset of parameters $(\theta_k)_{-\mathcal{U}}$, with the same dimensionality as $(\theta_{k'})_{-\mathcal{U}}$, is present in the current model k , we might choose a reversible jump proposal distribution which sets $(\theta_{k'})_{-\mathcal{U}} = (\theta_k)_{-\mathcal{U}}$ and proposes the remaining parameter vector $(\theta_{k'})_{\mathcal{U}}$ from its full conditional, $p((\theta_{k'})_{\mathcal{U}} | (\theta_{k'})_{-\mathcal{U}}, k', y)$. The reverse move would set $(\theta_k)_{-\mathcal{U}} = (\theta_{k'})_{-\mathcal{U}}$ and propose the remaining parameters in model k from their conditional $p((\theta_k)_{\mathcal{U}} | (\theta_k)_{-\mathcal{U}}, k', y)$.³ The reversible jump acceptance probability for such a move can then be derived as:

$$\alpha = \min \left(1, \frac{p(k' | (\theta_{k'})_{-\mathcal{U}} = (\theta_k)_{-\mathcal{U}}, y) q_1(k; k')}{p(k | (\theta_k)_{-\mathcal{U}}, y) q_1(k'; k)} \right) \quad (5)$$

where $p(k' | (\theta_{k'})_{-\mathcal{U}}, y) = \int_{(\Theta_{k'})_{\mathcal{U}}} p(k', (\theta_{k'})_{\mathcal{U}} | (\theta_{k'})_{-\mathcal{U}}, y) d(\theta_{k'})_{\mathcal{U}}$. A typical example where this might be used is the linear Gaussian model with conjugate priors, where the full conditional for the linear parameters is available. $(\theta_{k'})_{\mathcal{U}}$ might then be chosen to be the linear parameters for model k' , while $(\theta_{k'})_{-\mathcal{U}}$ could be the remaining unknown prior hyperparameters such as noise variances, etc., which are common to both models k and k' . These parameters, being of fixed dimensionality, can then be sampled in a separate step using a standard fixed-dimension MCMC method such as the Gibbs sampler or Metropolis-Hastings. Of course, a more sophisticated scheme might include in addition a random proposal to change the value of these ‘core’ parameters within the reversible jump proposal. In this way we can also deal with the case where the set of core parameters $(\theta_k)_{-\mathcal{U}}$ depends on k and hence the dimensionality of $(\theta_k)_{-\mathcal{U}}$ may also vary with k .

Note once again that the acceptance probability does not depend upon the sampled parameter values $(\theta_{k'})_{\mathcal{U}}$ or $(\theta_k)_{\mathcal{U}}$. In this case it depends upon the model proposal

³Note that in general $(\theta_k)_{\mathcal{U}}$ and $(\theta_{k'})_{\mathcal{U}}$ will be of differing dimensionality.

distributions and the posterior odds *conditional upon* $(\theta_k)_{-U} = (\theta_k)_{-U}$. In cases where $(\theta_k)_{-U}$ can be given a similar interpretation in both models (the parameters are ‘common’ to models k and k') this scheme is likely to yield a simple and effective model space sampler which takes advantage of any analytic structure within the model.

2.5 Example

To illustrate the principle of using partial analytic structure we examine a simple time series autoregression model uncertainty problem:

$$x_t = \sum_{i=1}^k a_i^{(k)} x_{t-i} + e_t, \quad e_t \stackrel{i.i.d.}{\sim} N(0, \sigma_e^2)$$

where $a^{(k)} = (a_i^{(k)}; i = 1, \dots, k)$ are the AR coefficients for a model of order k . For simplicity we side-step issues of stationarity and work with the conditional likelihood, which approximates the exact likelihood well for large N (Box et al., 1994):

$$p(x|a^{(k)}, \sigma_e^2, k) = \prod_{i=1}^N N\left(x_t - \sum_{i=1}^k a_i^{(k)} x_{t-i}\right), \quad x = [x_1 \dots x_N].$$

Conjugate normal-inverted Gamma priors are assumed for $a^{(k)}$ and σ_e^2 . A uniform prior is assumed for k over a range $1, \dots, k_{max}$, where k_{max} is set at 30 in this example (note that this upper limit is for computational convenience and is not required in general by the reversible jump scheme). Some partial analytic structure is then available in the form of the conditional distribution for $a^{(k)}$, $p(a^{(k)}|x, \sigma_e^2, k)$, which is multivariate Gaussian (Box et al., 1994). Thus we set $\theta_k = (a^{(k)}, \sigma_e^2)$, $(\theta_k)_U = a^{(k)}$ and $(\theta_k)_{-U} = \sigma_e^2$. The acceptance probability for model moves, following (5), is then:

$$\alpha = \min\left(1, \frac{p(k'|\sigma_e^2, x)q(k; k')}{p(k|\sigma_e^2, x)q(k'; k)}\right)$$

where $p(k|\sigma_e^2, x) = \int_{a^{(k)}} p(a^{(k)}, k|\sigma_e^2, x) da^{(k)}$, which is obtained analytically. σ_e^2 is updated at each iteration using a standard Gibbs sampling step.

1000 data points are simulated from an order 10 model with parameters arbitrarily chosen as

$$a^{(10)} = [0.9402, -0.4300, 0.4167, -0.4969, 0.4771, -0.5010, 0.0509, -0.2357, 0.4024, -0.1549]$$

and $\sigma_e^2 = 100$, as shown in figure 1. We chose a relatively large dataset to ensure that the conditional likelihood expression is accurate and also because this highlighted the differences between the two schemes considered. The schemes were:(1) the method

above based upon the partial analytic structure of the model and (2) a simple reversible jump implementation which proposes new parameters from an i.i.d. Gaussian, i.e.

$$\theta_{k'} = \begin{cases} [\sigma_e^2, a_1^{(k)} \dots a_{k'}^{(k)}], & k' \leq k, \\ [\theta_k, u_1 \dots u_{k'-k}], \quad u_i \stackrel{i.i.d.}{\sim} \text{N}(0, \sigma_u^2), & k' > k \end{cases}.$$

The acceptance probability for such a proposal is (see Green (1995, equation (8))), for $k' > k$:

$$\alpha = \min \left(1, \frac{p(a^{(k')}, k' | \sigma_e^2, x) q(k; k')}{p(a^{(k)}, k | \sigma_e^2, x) q(k'; k) \prod_{j=k+1}^{k'} \text{N}(\theta_j | 0, \sigma_u^2)} \right)$$

and the form of the fraction term is inverted for $k' < k$. We refer to this simple reversible jump implementation as the ‘stepwise’ sampler. In all other respects the two methods are identical, both including a ‘within model’ Gibbs move for $a^{(k)}$ and σ_e^2 at each iteration. The prior for $a^{(k)}$ was $\text{N}_k(0, 0.1I)$, and for σ_e^2 , $\text{IG}(10^{-5}, 10^{-5})$. The proposal distribution for the model orders, $q(\cdot; \cdot)$ has a discretised Laplacian shape, centred on the current model order. Note that this allows fairly frequent proposals to model orders which are distant from the current model.

The initial model order was assigned randomly from a uniform distribution over integers 1 to 30. The AR parameters were initialised to zero. The first step of the sampler was a Gibbs draw for σ_e^2 , so this does not require initialisation. Results for 30 runs of the partial analytic sampler are superimposed in figures 2 and 3, showing the consequences of randomly assigned initial model orders. We show only the initial hundred iterations as the sampler has always stabilised within the first few tens of iterations. By contrast we show the same results for the stepwise sampler under exactly the same initialisation conditions and proposing the new parameters from a zero mean normal distribution with variance $\sigma_u^2 = 0.1$. Note the different axis scaling for iteration number. The stepwise sampler has often not settled down within the first 1000 iterations and changes state relatively rarely. We do not claim that to have optimised the standard reversible jump implementation here as there are many possible options; however this comparison gives a reasonable flavour of the improvements which are achievable automatically, without any parameter tuning, simply through the use of the analytic structure of the model.

For full details of two variants on the approach used in this example see Troughton & Godsill (1997b).

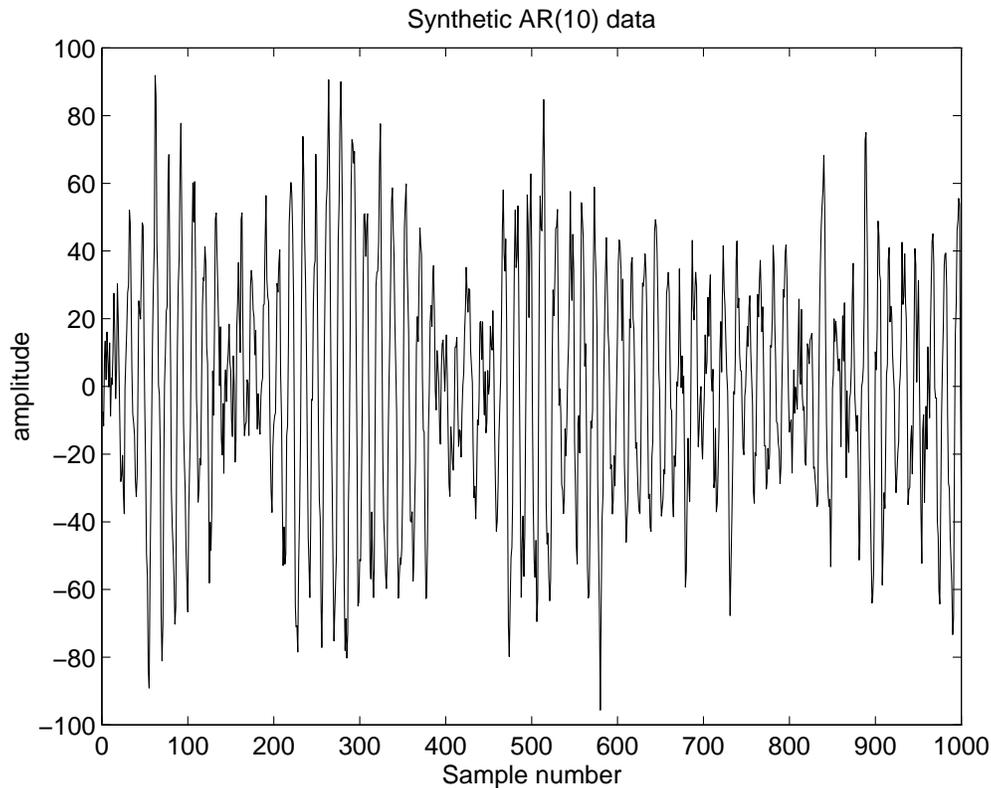


Figure 1: Synthetic AR(10) data

3 Discussion

The reversible jump scheme is a very effective way to apply a Metropolis-Hastings sampler to the composite space. One of its major advantages over the basic implementation of Carlin and Chib’s approach is that the values of parameters from models other than the two being compared at any given iteration (i.e. k and k') need not be updated or stored and pseudo-priors need not be devised. This is crucial for the large (or even infinite) sets of models which might need to be considered. We have demonstrated that there are close relationships between these methods and MCMC variable selection techniques for model space sampling, through consideration of a composite model space which encompasses all of the methods currently available. Simple analysis has shown that pseudo-priors (in the case of Carlin and Chib and MCMC variable selection) and parameter proposal distributions (in the case of reversible jump) which are designed to be close to the full posterior conditional for the parameters are likely to lead to very effective performance of both methods. Furthermore, we have proposed methods for taking advantage of partial analytic structure in a particular model to achieve efficient

model space moves.

The question arises as to whether new schemes can be derived based upon the composite model framework, which might be more effective in some scenarios than either Carlin and Chib's method or reversible jump, by applying some other MCMC scheme to the composite model. This question remains open and the topic of future work.

References

- BARBIERI, M. & O'HAGAN, A. (1996). A reversible jump MCMC sampler for Bayesian analysis of ARMA time series. Technical report, Università 'La Sapienza', Roma.
- BARNETT, G., KOHN, R. & SHEATHER, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Journal of Econometrics* **74**, 237–254.
- BESAG, J. (1989). A candidate's formula - a curious result in Bayesian prediction. *Biometrika* **76**, 183.
- BESAG, J. (1997). Discussion on: Bayesian analysis of mixtures with an unknown number of components by Richardson and Green. *Journal of the Royal Statistical Society, Series B* **59**, 731–758.
- BOX, G. E. P., JENKINS, G. M. & REINSEL, G. C. (1994). *Time Series Analysis, Forecasting and Control*. Prentice Hall, 3rd edition.
- CARLIN, B. P. & CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B* **57**, 473–484.
- CARTER, C. & KOHN, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83**, 589–601.
- CHIB, S. (1995). Marginal likelihood from the Gibbs output. *Journal of American Statistical Association* **90**, 1313–1321.
- CHIB, S. & GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.
- CLYDE, M., DESIMONE, H. & PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *Journal of American Statistical Association* **91**, 1197–1208.
- DELLAPORTAS, P., FORSTER, J. & NTZOUFRAS, I. (1997). On Bayesian model and variable selection using MCMC. Technical report, Dept. of Stats., Athens University of Econ. and Business.

- GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association* **85**, 398–409.
- GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**, 721–741.
- GEORGE, E. I. & MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *Journal of American Statistical Association* **88**, 881–889.
- GEWEKE, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics 5* pages 609–620.
- GODSILL, S. J. (1997). Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes. *International Statistical Review* **65**, 1–21.
- GODSILL, S. J. & RAYNER, P. J. W. (1996). Robust treatment of impulsive noise in speech and audio signals. In Berger, J., Betro, B., Moreno, E., Pericchi, L., Ruggeri, F., Salinetti, G. & Wasserman, L., editors, *Bayesian Robustness - proceedings of the workshop on Bayesian robustness, May 22-25, 1995, Rimini, Italy*, volume 29, pages 331–342. IMS Lecture Notes - Monograph Series.
- GREEN, P. & O'HAGAN, A. (1997). Carlin and Chib do not need to sample from pseudo-priors. Unpublished report.
- GREEN, P. J. (1995). Reversible jump Markov-chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- GRENDER, U. & MILLER, M. I. (1991). Jump-diffusion processes for abduction and recognition of biological shapes. Technical report, Electronic signals and systems research laboratory, Washington University.
- GRENDER, U. & MILLER, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B* **56**, 549–603.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- HOETING, J., RAFTERY, A. & MADIGAN, D. (1996). A method for simultaneous variable selection and outlier identification in linear-regression. *Computational statistics and data analysis* **22**, 251–270.
- HUERTA, G. & WEST, M. (1997). Priors and component structures in autoregressive time series models. *Duke University Research Paper* .
- KUO, L. & MALLICK, B. (1997). Variable selection for regression models. *Sankhya* (to appear).

- LIU, J. (1996). Metropolized Gibbs sampler: An improvement. Technical report, Stanford University.
- MADIGAN, D. & RAFTERY, A. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.
- MCCULLOCH, R. E. & TSAY, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis* **15**, 235–250.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.
- MORRIS, R. (1996). A sampling based approach to line scratch removal from motion picture frames. In *Proc. IEEE International Conference on Image Processing, Lausanne, Switzerland*.
- PHILLIPS, D. B. & SMITH, A. F. M. (1994). Bayesian model comparison via jump diffusions. Technical Report TR-94-20, Imperial College.
- RAFTERY, A. (1996). Hypothesis testing and model selection. In Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 163–187. Chapman and Hall.
- RAFTERY, A., MADIGAN, D. & HOETING, J. (1997). Bayesian model averaging for linear regression models. *Journal of American Statistical Association* **92**, 179–191.
- RICHARDSON, S. & GREEN, P. (1997). Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**, 731–758.
- STARK, J., FITZGERALD, W. & HLADKY, S. (1997). Multiple-order Markov chain Monte Carlo sampling methods with application to a changepoint model. Technical Report CUED/F-INFENG/TR.302, Department of Engineering, University of Cambridge.
- TROUGHTON, P. T. & GODSILL, S. J. (1997a). Bayesian model selection for time series using Markov chain Monte Carlo. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.
- TROUGHTON, P. T. & GODSILL, S. J. (1997b). A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. Technical Report CUED/F-INFENG/TR.304, Cambridge University Engineering Department.
- WEST, M. & HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*.

Springer.

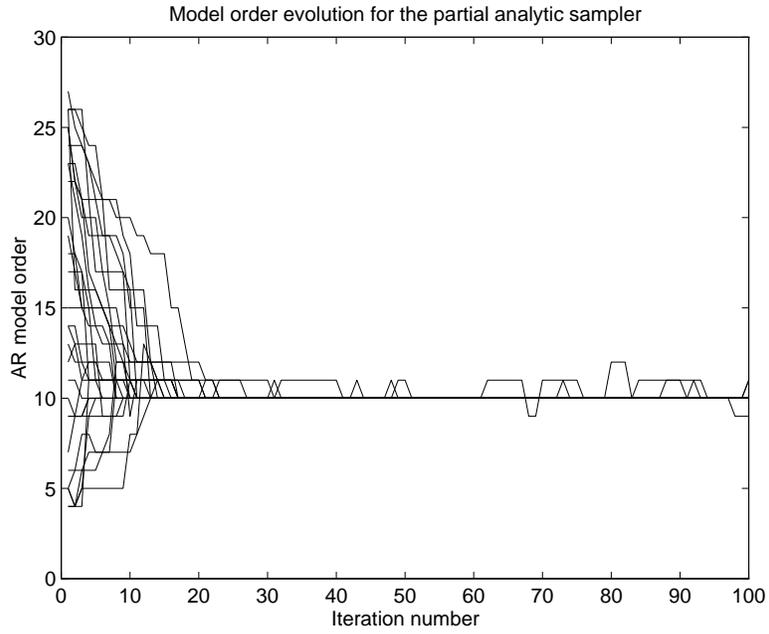


Figure 2: Model order evolution using the partial analytic sampler

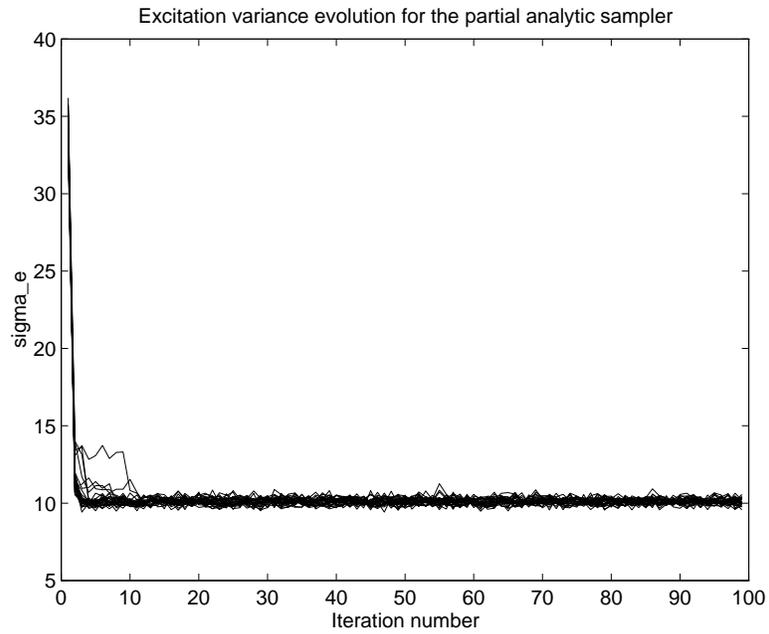


Figure 3: Evolution of σ_e using the partial analytic sampler

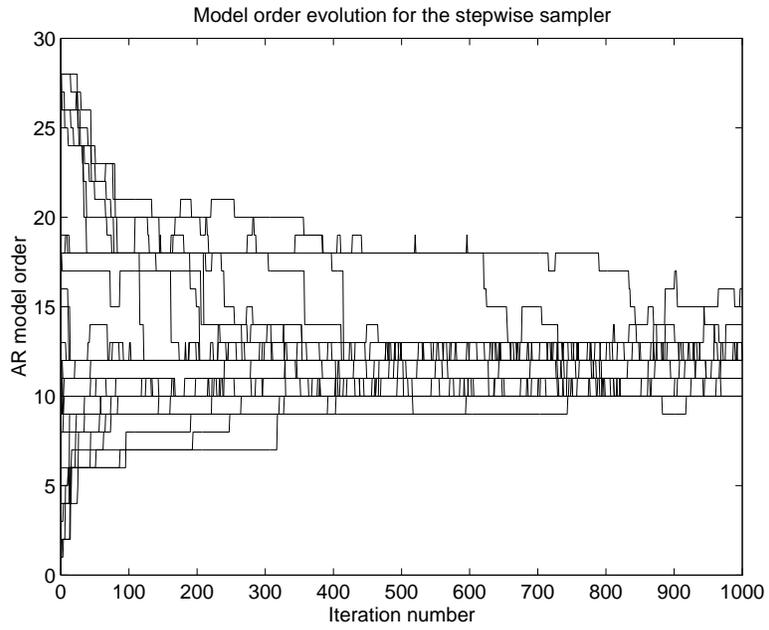


Figure 4: Model order evolution using the stepwise sampler

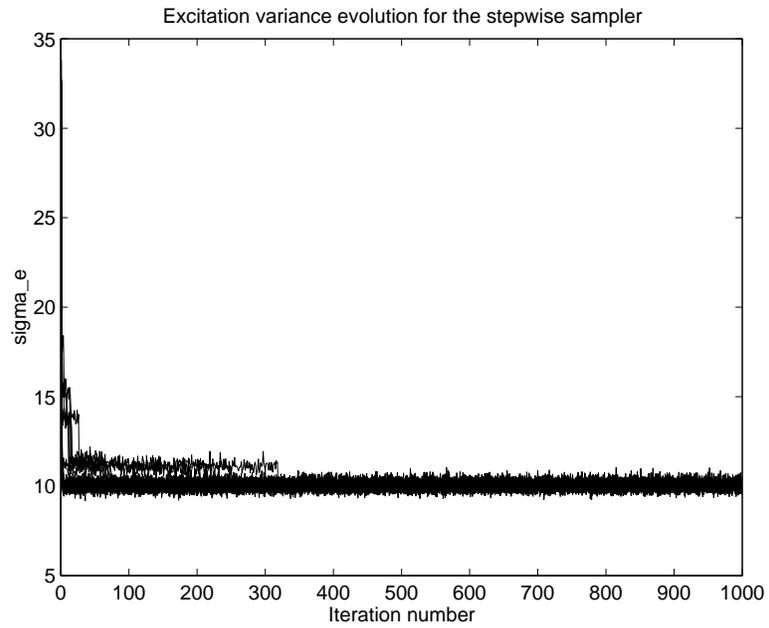


Figure 5: Evolution of σ_e using the stepwise sampler