1

Trans-dimensional Markov chain Monte Carlo

Peter J. Green
University of Bristol, UK

1 Introduction

Readers of this book will need no further convincing of the importance of Markov chain Monte Carlo (MCMC) in numerical calculations for highly structured stochastic systems, and in particular for posterior inference in Bayesian statistical models. Another chapter (Roberts, this volume) is devoted to discussion of some of the currently important research directions in MCMC generally. This chapter is more narrowly focussed on MCMC methods for what can be called 'trans-dimensional' problems, to borrow a nicely apt phrase from Roeder and Wasserman (1997): those where the dynamic variable of the simulation, the 'unknowns' in the Bayesian set-up, does not have fixed dimension.

Statistical problems where 'the number of things you don't know is one of the things you don't know' are ubiquitous in statistical modelling, both in traditional modelling situations such as variable selection in regression, and in more novel methodologies such as object recognition, signal processing, and Bayesian nonparametrics. All such problems can be formulated generically as a matter of joint inference about a model indicator k and a parameter vector θ_k, where the model indicator determines the dimension n_k of the parameter, but this dimension varies from model to model. Almost invariably in a frequentist setting, inference about these two kinds of unknown is based on different logical principles, but, at least formally, the Bayes paradigm offers the opportunity of a single logical framework — it is the joint posterior $p(k, \theta_k|Y)$ of model indicator and parameter given data Y that is the basis for inference. How can this be computed?

We set the joint inference problem naturally in the form of a simple Bayesian hierarchical model. We suppose given a prior $p(k)$ over models k in a countable set $\mathcal{K}$, and for each k, a prior distribution $p(\theta_k|k)$ and a likelihood $p(Y|k, \theta_k)$ for the data Y. For definiteness and simplicity of exposition, we suppose that $p(\theta_k|k)$ is a density with respect to n_k-dimensional Lebesgue measure, and that there are no other parameters, so that where there are parameters common to all models these are subsumed into each $\theta_k \in \mathcal{R}^{n_k}$. Additional parameters, perhaps in additional layers of a hierarchy, are easily dealt with. Note that in this chapter, all probability distributions are proper.

The joint posterior

$$p(k, \theta_k|Y) = \frac{p(k)p(\theta_k|k)p(Y|k,\theta_k)}{\sum_{k'\in\mathcal{K}} \int p(k')p(\theta'_{k'}|k')p(Y|k',\theta'_{k'})\mathrm{d}\theta'_{k'}}$$

can always be factorised as

$$p(k, \theta_k|Y) = p(k|Y)p(\theta_k|k, Y),$$

that is as the product of posterior model probabilities and model-specific parameter posteriors. This identity is very often the basis for reporting the inference, and in some of the methods mentioned below is also the basis for computation.

It is important to appreciate the generality of this basic formulation. In particular, note that it embraces not only genuine model-choice situations, where the variable k indexes the collection of discrete models under consideration, but also settings where there is really a single model, but one with a variable dimension parameter, for example a functional representation such as a series whose number of terms is not fixed. In the latter case, arising sometimes in Bayesian nonparametrics, for example, k is unlikely to be of direct inferential interest.

It can be argued that responsible adoption of a Bayesian hierarchical model of the kind introduced above presupposes that, for example, parameter priors $p(\theta_k|k)$ should be compatible in the sense that inference about functions of parameters that are meaningful in several models should be approximately invariant to k. Such compatibility could in principle be exploited in the construction of MCMC methods, although I am not aware of general methods for doing so. However, it is philosophically tenable that no such compatibility is present, and we shall not assume it.

Trans-dimensional MCMC has many applications other than to Bayesian statistics. Much of what follows will apply equally to them all; however, for simplicity, I shall use the Bayesian motivation and terminology throughout.

In Section 2, reversible jump MCMC is discussed, and this is related to other model-jumping approaches in Section 3. The following section treats alternatives to model-jumping, and Section 5 discusses and analyses some of the issues involved in choosing between the within- and across-model approaches. In Section 6, a simple fully-automated reversible jump sampler is introduced, and finally Section 7 notes some recent methodological extensions.

2 Reversible jump MCMC

In the direct approach to computation of the joint posterior $p(k, \theta_k|Y)$ via MCMC we construct a single Markov chain simulation, with states of the form (k, θ_k); we might call this an *across-model* simulation. We address other approaches in later sections.

The state space for such an across-model simulation is $\bigcup_{k\in\mathcal{K}}(\{k\} \times \mathcal{R}^{n_k})$; mathematically, this is not a particularly awkward object, and our construction involves no especially challenging novelties. However, such a state space is at least a little non-standard! Formally, our task is to construct a Markov chain on a general state space with a specified limiting distribution, and as usual in Bayesian MCMC for complex models, we use the Metropolis–Hastings paradigm to build a suitable reversible chain. As we see in the next subsection, on the face of it, this requires measure-theoretic notation, which may be unwelcome to some readers. The point of the 'reversible jump' framework is to render the measure theory invisible, by means of a construction using only ordinary densities. In fact, in the formulation given below, different and I hope improved from that of Green (1995), even the fact that we are jumping dimensions becomes essentially invisible!

2.1 Metropolis–Hastings on a general state space

We wish to construct a Markov chain on a state space $\mathcal{X}$ with invariant distribution π. As usual in MCMC we will consider only reversible chains, so the transition kernel P satisfies the detailed balance condition

$$\int_{(x,x')\in A\times B} \pi(\mathrm{d}x)P(x,\mathrm{d}x') = \int_{(x,x')\in A\times B} \pi(\mathrm{d}x')P(x',\mathrm{d}x) \qquad (2.1)$$

for all Borel sets $A, B \subset \mathcal{X}$. In Metropolis–Hastings, we make a transition by first drawing a candidate new state x' from the proposal measure $q(x, \mathrm{d}x')$ and then accepting it with probability $\alpha(x, x')$, to be derived below. If we reject, we stay in the current state, so that $P(x, \mathrm{d}x')$ has an atom at x. This contributes the same quantity $\int_{A\cap B} P(x, \{x\})\pi(\mathrm{d}x)$ to each side of (2.1); subtracting this leaves

$$\int_{(x,x')\in A\times B} \pi(\mathrm{d}x)q(x,\mathrm{d}x')\alpha(x,x') = \int_{(x,x')\in A\times B} \pi(\mathrm{d}x')q(x',\mathrm{d}x)\alpha(x',x). \quad (2.2)$$

It can be shown (Green 1995; Tierney 1998) that $\pi(\mathrm{d}x)q(x, \mathrm{d}x')$ is dominated by a symmetric measure μ on $\mathcal{X} \times \mathcal{X}$; let its density (Radon–Nikodym derivative) with respect to this μ be f. Then (2.2) becomes

$$\int_{(x,x')\in A\times B} \alpha(x,x')f(x,x')\mu(\mathrm{d}x,\mathrm{d}x') = \int_{(x,x')\in A\times B} \alpha(x',x)f(x',x)\mu(\mathrm{d}x',\mathrm{d}x)$$

and, using the symmetry of μ, this is clearly satisfied for all Borel A, B if

$$\alpha(x,x') = \min\left\{1, \frac{f(x',x)}{f(x,x')}\right\}.$$

This might be written more informally in the apparently familiar form

$$\alpha(x,x') = \min\left\{1, \frac{\pi(\mathrm{d}x')q(x',\mathrm{d}x)}{\pi(\mathrm{d}x)q(x,\mathrm{d}x')}\right\}. \qquad (2.3)$$

2.2 A constructive representation in terms of random numbers

Fortunately, the apparent abstraction in this prescription can be circumvented in most cases. By considering how the transition will be implemented in a computer program, the dominating measure and Radon–Nikodym derivatives can be generated implicitly. Take the case where $\mathcal{X} \subset \mathcal{R}^d$, and suppose π has a density (also denoted π) with respect to d-dimensional Lebesgue measure. At the current state x, we generate, say, r random numbers u from a known joint density g, and then form the proposed new state as some suitable deterministic function of the current state and the random numbers: $x' = h(x, u)$, say. The left-hand side of (2.2) can then be written as an integral with respect to (x, u):

$$\int_{(x,x')\in A\times B} \pi(x)g(u)\alpha(x,x')\,\mathrm{d}x\,\mathrm{d}u.$$

The reverse transition from x' to x would be made with the aid of random numbers $u' \sim g'$ giving $x = h'(x', u')$. If the transformation from (x, u) to (x', u') is a diffeomorphism (the transformation and its inverse are differentiable), then we can first write the right-hand side of (2.2) as an integral with respect to (x', u'), and then apply the standard change-of-variable formula. We then see that the $(d + r)$-dimensional integral equality (2.2) holds if

$$\pi(x)g(u)\alpha(x,x') = \pi(x')g'(u')\alpha(x',x)\left|\frac{\partial(x',u')}{\partial(x,u)}\right|,$$

where the last factor is the Jacobian of the diffeomorphism from (x, u) to (x', u'). Thus, a valid choice for α is

$$\alpha(x,x') = \min\left\{1, \frac{\pi(x')g'(u')}{\pi(x)g(u)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}, \tag{2.4}$$

involving only ordinary joint densities.

While this reversible jump formalism perhaps is a little indirect, it proves a flexible framework for constructing quite complex moves using only elementary calculus. In particular, the possibility that $r < d$ covers the case, typical in practice, that given $x \in \mathcal{X}$, only a lower-dimensional subset of $\mathcal{X}$ is reachable in one step. (The Gibbs sampler is the best-known example of this, since in that case only some of the components of the state vector are changed at a time, although the formulation here is more general as it allows the subset not to be parallel to the coordinate axes.) Separating the generation of the random innovation u and the calculation of the proposal value through the deterministic function $x' = h(x, u)$ is deliberate; it allows the proposal distribution $q(x, B) = \int_{x'\in B} h(x,u)g(u)\mathrm{d}u$ to be expressed in many different ways, for the convenience of the user.

2.3 The trans-dimensional case

However, the main benefit of this formalism is that expression (2.4) applies, without change, in a variable dimension context, if we use the same symbol $\pi(x)$

for the target density whatever the dimension of x in different parts of $\mathcal{X}$. Provided that the transformation from (x, u) to (x', u') remains a diffeomorphism, the individual dimensions of x and x' can be different. The dimension-jumping is indeed 'invisible'.

In this setting, suppose the dimensions of x, x', u and u' are d, d', r and r' respectively, then we have functions $h : \mathcal{R}^d \times \mathcal{R}^r \to \mathcal{R}^{d'}$ and $h' : \mathcal{R}^{d'} \times \mathcal{R}^{r'} \to \mathcal{R}^d$, used respectively in $x' = h(x, u)$ and $x = h'(x', u')$. For the transformation from (x, u) to (x', u') to be a diffeomorphism requires that $d + r = d' + r'$, so-called 'dimension-matching'; if this equality failed, the mapping and its inverse could not both be differentiable.

2.4 Details of application to the model-choice problem

Returning to our generic model-choice problem, we wish to use these reversible jump moves to sample the space $\mathcal{X} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{R}^{n_k})$ with invariant distribution π, which here is $p(k, \theta_k | Y)$.

Just as in ordinary MCMC, we typically need multiple types of moves to traverse the whole space $\mathcal{X}$. Each move is a transition kernel reversible with respect to π, but only in combination do we obtain an ergodic chain. The moves will be indexed by m in a countable set $\mathcal{M}$, and a particular move m proposes to take $x = (k, \theta_k)$ to $x' = (k', \theta'_{k'})$ or vice versa for a specific pair (k, k'); we denote $\{k, k'\}$ by $\mathcal{K}_m$. The detailed balance equation (2.2) is replaced by

$$\int_{(x,x') \in A \times B} \pi(\mathrm{d}x) q_m(x, \mathrm{d}x') \alpha_m(x, x') = \int_{(x,x') \in A \times B} \pi(\mathrm{d}x') q_m(x', \mathrm{d}x) \alpha_m(x', x)$$

for each m, where now $q_m(x, \mathrm{d}x')$ is the joint distribution of move type m and destination x'. The complete transition kernel is obtained by summing over m, so that for $x \notin B$, $P(x, B) = \sum_M \int_B q_m(x, \mathrm{d}x') \alpha_m(x, x')$, and it is easy to see that (2.1) is then satisfied.

The analysis leading to (2.3) and (2.4) is modified correspondingly, and yields

$$\alpha_m(x, x') = \min \left\{ 1, \frac{\pi(x')}{\pi(x)} \frac{j_m(x')}{j_m(x)} \frac{g'_m(u')}{g_m(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \right\}.$$

Here $j_m(x)$ is the probability of choosing move type m when at x, the variables x, x', u, u' are of dimensions d_m, d'_m, r_m, r'_m respectively, with $d_m + r_m = d'_m + r'_m$, we have $x' = h_m(x, u)$ and $x = h'_m(x', u')$, and the Jacobian has a form correspondingly depending on m.

Of course, when at $x = (k, \theta_k)$, only a limited number of moves m will typically be available, namely those for which $k \in \mathcal{K}_m$. With probability $1 - \sum_{m:k \in \mathcal{K}_m} j_m(x)$ no move is attempted.

2.5 Some remarks and ramifications

In understanding the reversible jump framework, it may be helpful to stress the key role played by the joint state-proposal equilibrium distributions. The fact that the degrees of freedom in these joint distributions are unchanged when x and

x' are interchanged allows the possibility of reversible jumps across dimensions, and these distributions directly determine the move acceptance probabilities.

Note that the framework gives insights into Metropolis–Hastings that apply quite generally. State-dependent mixing over a family of transition kernels in general infringes detailed balance, but is permissible if, as here, the move probabilities $j_m(x)$ enter properly into the acceptance probability calculation. Note also the contrast between this randomised proposal mechanism, and the related idea of mixture proposals, where the acceptance probability does not depend on the move actually chosen; see the discussion in Besag *et al.* (1995, appendix 1). Contrary to some accounts that connect it with the jump in dimension, the Jacobian comes into the acceptance probability simply through the fact that the proposal destination $x' = h(x, u)$ is specified indirectly.

Finally, note that in a large class of problems involving nested models, the only dimension change necessary is the addition or deletion of a component of the parameter vector (think of polynomial regression, or autoregression of variable order). In such cases, omission of a component is often equivalent to setting a parameter to zero. These problems can be handled in a seemingly more elementary way, through allowing proposal distributions with an atom at zero: the usual Metropolis–Hastings formula for the acceptance probability holds for densities with respect to arbitrary dominating measures, so the reversible jump formalism is not explicitly needed. Nevertheless, it leads to exactly the same algorithm.

Other authors have provided different pedagogical descriptions of reversible jump. Waagepetersen and Sorensen (2001) provide a tutorial following the lines of Green (1995) but in much more detail, and Besag (1997, 2000) gives a novel formulation in which variable dimension notation is circumvented by embedding all θ_k within one compound vector; this has something in common with the product-space formulations in the next subsection.

3 Relations to other across-model approaches

Several alternative formalisms for across-model simulation are more or less closely related to reversible jump.

Jump diffusion. In addressing challenging computer vision applications, Grenander and Miller (1994) proposed a sampling strategy they termed jump diffusion. This comprised two kinds of move — between-model jumps, and within-model diffusion according to a Langevin stochastic differential equation. Since in practice, continuous-time diffusion has to be approximated by a discrete-time simulation, they were in fact using a trans-dimensional Markov chain. Had they corrected for the time discretisation by a Metropolis–Hastings accept/reject decision (giving a so-called Metropolis-adjusted Langevin algorithm or MALA) (Besag 1994), this would have been an example of reversible jump.

Phillips and Smith (1996) applied jump-diffusion creatively to a variety of Bayesian statistical tasks, including mixture analysis, object recognition and variable selection.

Point processes, with and without marks. Point processes form a natural example of a distribution with variable-dimension support, since the number of points in view is random; in the basic case, a point has only a location, but more generally may be accompanied by a *mark*, a random variable in a general space.

A continuous time Markov chain approach to simulating certain spatial point processes, by regarding them as the invariant distributions of spatial birth-and-death processes, was suggested and investigated by Preston (1977) and Ripley (1977). More recently, Geyer and Møller (1994) proposed a Metropolis–Hastings sampler, as an alternative to using birth-and-death processes; their construction is a special case of reversible jump.

Stephens (2000) notes that various trans-dimensional statistical problems can be viewed as abstract marked point processes: in these models, the items of which there are a variable number are regarded as marked points. For example in a normal mixture model the points represent the mean–variance pairs of the components, marked with the component weights. Stephens borrows the birth-and-death simulation idea to develop a methodology for finite mixture analysis, and also suggests that the approach appears to have much wider application, citing change point analysis and regression variable selection as partially worked examples. The key feature of these three settings that allows the approach to work is the practicability of integrating out latent variables so that the likelihood is fully available. See also Hurn *et al.* (2001) for application to mixtures of regressions. Cappé *et al.* (2001) have recently given a rather complete analysis of the relationship between reversible jump and continuous time birth-and-death samplers.

Product-space formulations. Several relatives of reversible jump work in a product space framework, that is, one in which the simulation keeps track of all θ_k, not only the 'current' one. The state space is therefore $\mathcal{K} \times \otimes_{k\in\mathcal{K}}\mathcal{R}^{n_k}$ instead of $\bigcup_{k\in\mathcal{K}}(\{k\}\times\mathcal{R}^{n_k})$. This has the advantage of circumventing the trans-dimensional character of the problem, at the price of requiring that the target distribution be augmented to model all θ_k simultaneously. For some variants of this approach, this is just a formal device, for others it leads to significantly extra work.

Let θ_{-k} denote the composite vector consisting of all $\theta_l, l \neq k$ catenated together. Then the joint distribution of $(k, (\theta_l : l \in \mathcal{K}), Y)$ can be expressed as

$$p(k)p(\theta_k|k)p(\theta_{-k}|k,\theta_k)p(Y|k,\theta_k), \tag{3.1}$$

since we make the natural assumption that $p(Y|k,(\theta_l : l \in \mathcal{K})) = p(Y|k,\theta_k)$. It is easily seen that the third factor $p(\theta_{-k}|k,\theta_k)$ has no effect on the joint posterior $p(k,\theta_k|Y)$; the choice of these conditional distributions, which Carlin and Chib (1995) call 'pseudo-priors', is entirely a matter of convenience, but may influence the efficiency of the resulting sampler.

Carlin and Chib (1995) adopted pseudo-priors that were conditionally independent: $p(\theta_{-k}|k,\theta_k) = \prod_{l \neq k} p(\theta_l|k)$, and assumed $p(\theta_l|k)$ does not depend on k for $k \neq l$. They used a Gibbs sampler, updating k and all θ_l in turn. This evidently involves sampling from the pseudo-priors, and they therefore propose to design these pseudo-priors to ensure reasonable efficiency, which requires their approximate matching to the posteriors: $p(\theta_l|k) \approx p(\theta_l|l,Y)$.

Green and O'Hagan (1998) pointed out both that Metropolis–Hastings moves could be made in this setting, and that in any case there was no need to update $\{\theta_l, l \neq k\}$ to obtain an irreducible sampler. In this form the pseudo-priors are only used in computing the update of k. Dellaportas *et al.* (2002) proposed and investigated a 'Metropolised Carlin and Chib' approach, in which joint model indicator/parameter updates were made, and in which it is only necessary to resample the parameter vectors for the current and proposed models.

Godsill (2001) introduces a general 'composite model space' framework that embraces all of these methods, including reversible jump, facilitating comparisons between them. He devised the formulation (3.1), or rather, a more general version in which the parameter vectors θ_k are allowed to overlap arbitrarily, each θ_k being identified with a particular sub-vector of one compound parameter. This framework helps to reveal that a product-space sampler may or may not entail possibly cumbersome additional simulation, updating parameters that are not part of the 'current' model. It also gives useful insight into some of the important factors governing the performance of reversible jump, and Godsill offers some suggestions on proposal design.

Godsill's formulation deserves further attention, as it provides a useful language for comparing approaches, and in particular examining one of the central unanswered questions in trans-dimensional MCMC. Suppose the simulation leaves model k and later returns to it. With reversible jump, the values of θ_k are lost as soon as we leave k, while with some versions of the product-space approach, the values are retained until k is next visited. Intuitively either strategy has advantages and disadvantages for sampler performance, so which is to be preferred?

4 Alternatives to joint model-parameter sampling

The direct approach of a single across-model simulation is in many ways the most appealing, but alternative indirect methods that treat the unknowns k and θ_k differently should not be neglected.

Integrating out the parameters. If in each model k, the prior is conjugate for the likelihood, then $p(\theta_k|k,Y)$ may be explicitly available, and thence can be calculated the *marginal likelihoods*

$$p(Y|k) = \frac{p(\theta_k|k)p(Y|k,\theta_k)}{p(\theta_k|k,Y)}$$

and finally the posterior probabilities $p(k|Y) \propto p(k)p(Y|k)$. In the very limited cases where this is possible, Bayesian inference about k, and about θ_k given k, can be conducted separately, and trans-dimensional simulations are not needed.

The approach has been taken a little further by Godsill (2001), who considers cases of 'partial analytic structure', where some of the parameters in θ_k may be integrated out, and the others left unchanged in the move that updates the model, to give an across-model sampler with probable superior performance.

Within-model simulation. If samplers for the within-model posteriors $p(\theta_k|Y, k)$ are available for each k, then joint posterior inference for (k, θ_k) can be constructed by combining separate simulations conducted within each model. See Carlin and Louis (1996, §6.3.1) for more detailed discussion.

The posterior $p(\theta_k|Y, k)$ for the parameters θ_k is in any case a within-model notion, and is the target for an ordinary Bayesian MCMC calculation for model k. Since

$$\frac{p(k_1|Y)}{p(k_0|Y)} = \frac{p(k_1)}{p(k_0)} \frac{p(Y|k_1)}{p(Y|k_0)}$$

(the second factor being the *Bayes factor* for model k_1 vs. k_0), to find the posterior model probabilities $p(k|Y)$ for all k it is sufficient to estimate the marginal likelihoods

$$p(Y|k) = \int p(\theta_k, Y|k)\, \mathrm{d}\theta_k$$

separately for each k, using individual MCMC runs. Several different methods have been devised for this task.

Noting that $p(Y|k)$ can be expressed as $\{\int [p(\theta_k|k, Y)/p(Y|k, \theta_k)]\, \mathrm{d}\theta_k\}^{-1}$ or more directly as $\int p(Y|k, \theta_k)p(\theta_k|k)\, \mathrm{d}\theta_k$, leads respectively to the estimates

$$\widehat{p}_1(Y|k) = N \Bigg/ \sum_{t=1}^{N} \left\{ p(Y|k, \theta_k^{(t)}) \right\}^{-1} \quad \text{and} \quad \widehat{p}_2(Y|k) = N^{-1} \sum_{t=1}^{N} p(Y|k, \theta_k^{(t)}),$$

based on MCMC samples $\theta_k^{(1)}, \theta_k^{(2)}, \ldots$ from the posterior $p(\theta_k|Y, k)$ and the prior $p(\theta_k|k)$, respectively. Both of these are simulation-consistent, but have high variance, with possibly few terms contributing substantially to the sums in each case. Composite estimates, based like $\widehat{p}_1$ and $\widehat{p}_2$ on the importance sampling identity $E_p(f) = E_q(fp/q)$, perform better, including those of Newton and Raftery (1994) and Gelfand and Dey (1994). For example, Newton and Raftery propose to simulate from a mixture $\widetilde{p}(\theta_k; Y, k)$ of the prior and posterior, and use

$$\widehat{p}_3(Y|k) = \frac{\sum_{t=1}^{N} p(Y|k, \theta_k^{(t)}) w(\theta_k^{(t)})}{\sum_{t=1}^{N} w(\theta_k^{(t)})},$$

where $w(\theta_k) = p(\theta_k|k)/\widetilde{p}(\theta_k; Y, k)$.

Chib (1995) has introduced new, indirect, estimates of the marginal likelihood based on the identity $p(Y|k) = p(Y|k, \theta_k^\star)p(\theta_k^\star|k)/p(\theta_k^\star|k, Y)$ for any fixed parameter point $\theta_k^\star$. The factors in the numerator are available, and in contexts where the parameter can be decomposed into blocks with explicit full conditionals, the denominator can be estimated using simulation calculations that use the same Gibbs sampling steps as the posterior simulation. Note, however, that Neal (1999) has demonstrated that Chib's application of this idea to mixture models is incorrect. Chib and Jeliazkov (2001) extend the idea to cases where Metropolis–Hastings is needed.

5 Some issues in choosing a sampling strategy

Several studies have addressed the strengths and weaknesses of reversible jump MCMC and the other trans-dimensional setups above compared to within-model simulations that compute marginal likelihoods and thence Bayes factors. Particularly noteworthy are Dellaportas *et al.* (2002), Godsill (2001) and Han and Carlin (2001). Each of these discusses some of the issues involved and provides comparisons of implementations and performance on test problems, although, understandably in the present state of our knowledge with these methods, it is hard to see any of these as entirely definitive.

One of the key matters influencing the choice here is the number of models to be entertained, taking account of the degree of homogeneity between them. The ideal situation for the 'within-model' strategy would be a case where the models are all of a different character, and fully-tested samplers with acceptable performance are already available for each. In such a case, building an across-model sampler could be very laborious compared to adding marginal likelihood calculations to each model separately.

Some authors have recorded poor performance with reversible jump methods. Since reversible jump algorithms embrace all Metropolis–Hastings methods for the across-model state space, it is hard to believe that there are no methods in this huge class that would give acceptable performance. It would be fairer to say that existing examples of reversible jump implementations may be poor templates for constructing samplers in some new situations. A difficulty is that the across-model state space may be hard to visualise so that some of the intuition that guides construction of samplers in simpler spaces is not available.

Others have deemed reversible jump methods cumbersome to construct and difficult to tune. There seems to be a need for further methodological work, developing broader classes of across-model samplers, with associated visualisation techniques, to assist in construction and tuning. Very recent work by Brooks *et al.* (2000) may be a good step in this direction; see Section 7.2. Of course, as in other domains for MCMC, fully-automated sampler construction would be a tremendous advantage: a very limited step towards this is introduced in Section 6 below.

Finally, the across-model approach does have another potential benefit — the possibility that jumping models can improve mixing. This is discussed next.

5.1 Is it good to jump?

There are various not entirely substantiated claims in the literature to the effect that jumping between parameter subspaces is either inherently damaging to MCMC performance and should therefore be avoided where possible, or alternatively that it is helpful for performance, and might even be attempted when it is not strictly necessary.

For example, Richardson and Green (1997) describe a simple experiment, illustrated in their Fig. 9, demonstrating that in a particular example of a mixture problem with a strongly multimodal posterior, mixing is clearly improved by using a trans-dimensional sampler, while Han and Carlin (2001) claim to have 'intuition that some gain in precision should accrue to MCMC methods that avoid a model space search'.

In truth, the proper answer is 'it depends', but some simple analysis does reveal some of the issues. There are three main situations that might be considered: in the first, we require full posterior inference about (k, θ_k). A second possibility is that we wish to make within-model inference about θ_k separately, for each of a (perhaps small) set of values of k. The third case is where k is really fixed, and the other models are ruled out a priori. This third option is clearly the least favourable for trans-dimensional samplers: visits of the (k, θ_k) chain to the 'wrong' models are wasted from the point of view of extracting useful posterior information; let us try to analyse whether superior mixing in these other models can nevertheless make it worthwhile to use a trans-dimensional sampler.

5.2 The two-model case

For simplicity, we suppose there are just two models, $k = 1$ and 2, and let π_k denote the distribution of θ_k given k: only π_1 is of interest. We have transition kernels Q_{11}, Q_{22}, with $\pi_k Q_{kk} = \pi_k$ for each k; (we use a notation apparently aimed at the finite state space case, but it is quite general: for example, πQ means the probability measure $(\pi Q)(B) = \int \pi(\mathrm{d}x)Q(x, B)$). We now consider the option of also allowing between-model transitions, with the aid of kernels Q_{12} and Q_{21}; for realism, these are improper distributions, integrating to less than 1, reflecting the fact that in practice across-model Metropolis–Hastings moves are frequently rejected. When a move is rejected, the chain does not move, contributing a term to the 'diagonal' of the transition kernel; thus we suppose there exist diagonal kernels D_1 and D_2, and we have the global balance conditions for the across-model moves: $\pi_1 D_1 + \pi_2 Q_{21} = \pi_1$ and $\pi_2 D_2 + \pi_1 Q_{12} = \pi_2$.

Assuming that we make a random choice between the two moves available from each state, α and β being the probabilities of choosing to attempt the between-model move in models 1, 2 respectively, the overall transition kernel for the across-model sampler is

$$P = \begin{pmatrix} (1-\alpha)Q_{11} + \alpha D_1 & \alpha Q_{12} \\ \beta Q_{21} & (1-\beta)Q_{22} + \beta D_2 \end{pmatrix}$$

using an obvious matrix notation. The invariant distribution is easily seen to be

$\pi = (\gamma\pi_1, (1-\gamma)\pi_2)$, where $\gamma = \beta/(\alpha+\beta)$.

Now suppose we run the Markov chain given by P, but look at the state only when in model 1. By standard Markov chain theory, the resulting chain has kernel $\widetilde{Q}_{11} = (1-\alpha)Q_{11} + \alpha D_1 + \alpha Q_{12}\{I - (1-\beta)Q_{22} - \beta D_2\}^{-1}\beta Q_{21}$. The comparison we seek is that between using Q_{11} or the more complicated strategy that amounts to using $\widetilde{Q}_{11}$, but we must take into account differences in costs of computing. Suppose that executing Q_{11} or Q_{22} has unit cost per transition, while attempting and executing the across-model moves has cost c times greater. Then, per transition, the equilibrium cost of using P is $\gamma(1-\alpha) + \gamma\alpha c + (1-\gamma)\beta c + (1-\gamma)(1-\beta)$, and this gives on average γ visits to model 1. The relative cost in computing resources of using $\widetilde{Q}_{11}$ instead of Q_{11} therefore simplifies to $(1-\alpha) + 2\alpha c + \alpha(1-\beta)/\beta$ (using the relationship $\gamma\alpha = (1-\gamma)\beta$).

If we choose to measure performance by asymptotic variance of a specific ergodic average, then we have integrated autocorrelation times τ and $\widetilde{\tau}$ for Q_{11} and $\widetilde{Q}_{11}$ respectively, and jumping models is a good idea if

$$\tau < \widetilde{\tau}\{(1-\alpha) + 2\alpha c + \alpha(1-\beta)/\beta\}.$$

Of course, $\widetilde{\tau}$ depends on α and β.

5.3 Finite state space example

It is interesting to compute these terms for toy finite-state-space examples where the eigenvalue calculations can be made explicitly. For example, taking $D_1 = D_2 = 0.8I$, corresponding to an 80% rejection rate for between-model moves, and all the Q matrices to be symmetric reflecting random walks on $m = 10$ states, with differing probabilities of moving, to model differently 'sticky' samplers, specifically $(Q_{11})_{i,i\pm1} = 0.03$, $(Q_{12})_{i,i\pm1} = 0.2 \times 0.1$, $(Q_{21})_{i,i\pm1} = 0.2 \times 0.1$, and $(Q_{22})_{i,i\pm1} = 0.3$, we find that model jumping is worthwhile for all c up to about 15, with optimal $\alpha \approx 1$ and $\beta \approx 0.1$. This is a situation where the rapid mixing in model 2 compared to that in model 1 justifies the expense of jumping from 1 to 2 and back again.

5.4 Tempering-by-embedding

Such considerations raise the possibility of artificially embedding a given statistical model into a family indexed by k, and conducting an across-model simulation simply to improve performance — that is, as a kind of simulated tempering (Marinari and Parisi 1992). A particular example of the benefit of doing so was given by Hodgson (1999) in constructing a sampler for restoration of ion channel signals. A straightforward approach to this task gave poor mixing, essentially because of high posterior correlation between the model hyperparameters and the hidden binary signal. This correlation is higher when the data sequence is longer, so a tempering-by-embedding solution was to break the data into blocks, with the model hyperparameters allowed to change between adjacent blocks. The part of the prior controlling this artificial model elaboration was adjusted empirically to give moderately high rates of visiting the real model, while spending sufficient time in the artificial heterogeneous models for the harmful correlation to be substantially diluted.

Further evidence that model-jumping can provide effective tempering, admittedly in a somewhat contrived setting, was provided by Richardson and Green (1997). They compared fixed-k and variable-k samplers for a normal mixture problem with k components, applied to a symmetrised bimodal data set. In this case, there was substantial posterior support for $k = 2$ and 4; MCMC-based inference about parameters conditional on $k = 3$ was greatly superior using the variable-k sampler.

6 An automatic generic trans-dimensional sampler

The possibility of automating the construction of a MCMC sampler for any given target distribution is attractive but elusive. It would be a tremendous practical advantage if the user could just specify the target in algebraic form, perhaps together with a few numerical constants such as starting values, and leave the computer both to construct an algorithm and then run it to create a reliable sample.

The nearest we can come to this ideal at present, for sampling from a fixed-dimensional density, is the random-walk Metropolis (RWM) sampler (see Roberts, this volume), in the most simple form where all variables are simultaneously updated. Other possibilities, requiring a little more user input, are Langevin methods, or the hybrid samplers of Duane *et al.* (1987). RWM is not a panacea. From a theoretical perspective, it is imperfect since even geometric ergodicity is not guaranteed, as it requires conditions on the relative size of the tails of the target and proposal densities. In fact, no kind of ergodicity is certain, since there may be holes in the support of the target and/or the proposal density which could prevent irreducibility, but such pathologies are easily avoided. There is also the important practical consideration that updating all variables at once prevents the exploitation of factorisations of the target that make the acceptance probabilities for lower-dimensional updates particularly cheap to compute.

In spite of these drawbacks, the RWM methods are useful and it would be valuable to have an analogous class of methods for trans-dimensional problems, particularly for exploratory use. In this section, we propose a rather naive approach to this quest, but as experiments show, the results are quite promising.

Suppose that for each model k, we are given a fixed n_k-vector μ_k and a fixed $n_k \times n_k$-matrix B_k. Consider the situation where we are currently in state (k, θ_k) and have proposed a move to model k', drawn from some transition matrix $(r_{k,k'})$. The form of the proposed new parameter vector depends on whether $n_{k'}$ is less than, equal to, or more than n_k. We set:

$$\theta'_{k'} = \begin{cases} \mu_{k'} + B_{k'}[RB_k^{-1}(\theta_k - \mu_k)]_1^{n_{k'}} & \text{if } n_{k'} < n_k \\ \mu_{k'} + B_{k'}RB_k^{-1}(\theta_k - \mu_k) & \text{if } n_{k'} = n_k \\ \mu_{k'} + B_{k'}R\begin{pmatrix} B_k^{-1}(\theta_k - \mu_k) \\ u \end{pmatrix} & \text{if } n_{k'} > n_k \end{cases} .$$

Here $[\cdots]_1^m$ denotes the first m components of a vector, R is a fixed orthogonal matrix of order $\max\{n_k, n_{k'}\}$, and u is a $(n_{k'} - n_k)$-vector of random numbers with density $g(u)$.

Note that if $n_{k'} \leq n_k$, the proposal is deterministic (apart from the choice of k'). Since everything is linear, the Jacobian is trivially calculated: if $n_{k'} > n_k$, we have

$$\left| \frac{\partial(\theta_{k'})}{\partial(\theta_k, u)} \right| = \frac{|B_{k'}|}{|B_k|}.$$

Thus the acceptance probability is $\min\{1, A\}$, where

$$A = \frac{p(k', \theta'_{k'}|y)}{p(k, \theta|y)} \frac{r_{k',k}}{r_{k,k'}} \frac{|B_{k'}|}{|B_k|} \times \begin{cases} g(u) & \text{if } n_{k'} < n_k \\ 1 & \text{if } n_{k'} = n_k \\ g(u)^{-1} & \text{if } n_{k'} > n_k \end{cases}.$$

Since it is orthogonal, the matrix R plays no role in this calculation.

If the model-specific targets $p(\theta_k|k, y)$ were normal distributions, with means μ_k and variances $B_k B_k^T$, if the innovation variables u were standard normal, and if we could choose $r_{k,k'}/r_{k',k} = p(k'|Y)/p(k|Y)$, these proposals would already be in detailed balance, with no need to compute the Metropolis–Hastings accept/reject decision. This is the motivation for the idea.

This suggests that, providing the $p(\theta_k|k, y)$ are reasonably unimodal, with mean and variance approximately equal to μ_k and $B_k B_k^T$, this simple sampler may be effective. A simple modification, likely to give performance more robust to heavy tails in the targets, would be to use t-distributions in place of the normals for u. Another modification, plausibly likely on general grounds to improve mixing, is to randomise over the orthogonal matrix R, or, more simply, take R to be a random permutation matrix. By the usual argument about randomised proposals (Besag *et al.* 1995, Appendix 1), this randomisation can be ignored when calculating the acceptance probability.

In applications, we are only likely to have approximations to the mean and variances of $p(\theta_k|k, y)$ when we can conduct pilot runs within each model separately — thus limiting the idea to cases where the set of models $\mathcal{K}$ is finite and small. In our implementation, we loop over these models and perform a short run of RWM on each to estimate the means μ_k and variances $B_k B_k^T$. We then find the lower triangular square root of the variance B_k (and its determinant) by Cholesky decomposition; the advantage of using a lower triangular B_k is that we can use forward substitution to multiply B_k^{-1} into a vector.

Finally, the idea might have broader applicability if the pilot runs were used also to detect and correct gross departures from normality — perhaps a transformation to reduce skewness could be estimated, for example. We have not explored such modifications.

6.1 Examples

This method has been implemented as a stand-alone Fortran program, available from the author by email (`P.J.Green@bris.ac.uk`), which calls a function written by the user to compute $\log p(k, \theta_k, y)$. The only other information required about the problem, also provided by this function, are the number of models, their dimensions, and rough settings for the centre and spread of each variable, used for initial values and spread parameters for the RWM moves. The code is set up to alternate between model-jumping moves as described above, and within-model moves by RWM.

We have tried the approach on two non-trivial examples; the codes for these two were identical apart from the information just described.

(a) Variable selection in a small logistic regression problem. Dellaportas *et al.* (2002) illustrate their comparisons between model-jumping algorithms on a small data set. This is a 2×2 factorial experiment with a binomially distributed response variable. All 5 interpretable models are entertained, with numbers of parameters (n_k) equal to 1, 2, 2, 3 and 4 respectively. We follow Dellaportas *et al.* exactly in terms of prior settings, etc. One million sweeps of the automatic sampler, many more than is needed for reliable results, takes about 18 seconds on a 800MHz PC. The acceptance rate for the model-jumping moves was 29.4%, and the integrated autocorrelation time for estimating $E(k|y)$ was estimated by Sokal's method (Green and Han 1992) to be 2.90. The posterior model probabilities were computed to be (0.005, 0.493, 0.011, 0.439, 0.052), consistent with the results of Dellaportas *et al.*

(b) Change point analysis for a point process. We revisit the change point analysis of the coal mine disaster data used in Green (1995). In this illustration, we condition on the number of change points k lying in the set $\{1, 2, 3, 4, 5, 6\}$ (which covers most of the posterior probability). All the prior settings, etc., are as in Green (1995). There are $2k + 1$ parameters in model k. For this problem, 1 million sweeps takes about 28 seconds on a 800MHz PC. The posterior for the number of change points was estimated to be (0.058, 0.251, 0.294, 0.236, 0.117, 0.044) for the values $k = 1$ to 6. Note that this differs somewhat from the results reported in Green (1995); in fact if the sampler derived there is run for 200 000 sweeps instead of 40 000, the results become very similar. On this problem, the automatic sampler mixes much less well: the acceptance rate for model-jumping is 5.9%, while the integrated autocorrelation time estimate rises to 118. This decline in performance is presumably due to the extreme multi-modal character of many of the parameter posteriors.

For comparison, the sampler described in Green (1995) takes 14 seconds for 1 000 000 sweeps on this computer, with an acceptance rate of 21% and estimated autocorrelation time of 67.8. On this basis, the relative efficiency of the automatic sampler is only $(14 \times 67.8)/(28 \times 118) \approx 29\%$, but of course the implementation time was far less.

6.2 Limitations of this approach

I have stressed that this automatic sampler cannot be expected to have very broad applicability. However, its successful use on the second example above shows that it can be surprisingly tolerant to multimodality in the model-specific targets. (As seen in Figs. 3 and 4 of Green (1995), multimodality is evident even for $k = 2$, and it rapidly becomes very severe for larger k.) In such cases, it is necessary for the proposal spread factors provided to be sufficiently large that there is adequate jumping between modes, in the pilot runs within each model.

This approach is unlikely to be useful for more than a small set of models, so that, for example, variable selection between many variables is probably out of reach. It may, however, be worth exploring whether quite crude approximations to the means and variances of each target give adequate performance, and whether such approximations can be generated for variable selection problems without conducting pilot runs on all models.

7 Methodological extensions

7.1 Delayed rejection

An interesting modification to Metropolis–Hastings is the splitting rejection idea of Tierney and Mira (1999), which has recently been extended to the reversible jump setting by Green and Mira (2001), who call it delayed rejection.

The idea is simple: if a proposal is rejected, instead of 'giving up', staying in the current state, and advancing time to the next transition, we can instead attempt a second proposal, usually from a different distribution, and possibly dependent on the value of the rejected proposal. It is possible to set the acceptance probability for this second-stage proposal so that detailed balance is obtained, individually within each stage. The idea can be extended to further stages.

By the results of Peskun (1973), generalised in Tierney (1998), such a strategy is always advantageous in terms of reducing asymptotic variances of ergodic averages, on a sweep-by-sweep basis, since the probability of moving increases by stage. Whether it is actually worth doing will depend on whether the reduction in Monte Carlo variance compensates for the additional computing time for the extra stages; the experiments reported in Green and Mira (2001) suggest that this can be the case.

The second-stage acceptance probability is calculated by an argument along the same lines as that in Section 2 above. We use two vectors of random numbers u_1 and u_2, drawn from densities g_1 and g_2 respectively, and two deterministic functions mapping these and the current state into the proposed new states, $y = h_1(x, u_1)$ and $z = h_2(x, u_1, u_2)$, respectively. Both u_1 and u_2 appear in the expression for z to allow this second-stage proposal to be dependent on the rejected first-stage candidate y; for example, z may be a move in a different 'direction' in some sense.

The first-stage proposal is accepted with probability $\alpha_1(x, y)$ calculated as usual:

$$\alpha_1(x, y) = \min\left\{1, \frac{\pi(y)g_1'(u_1')}{\pi(x)g_1(u_1)}\left|\frac{\partial(y, u_1')}{\partial(x, u_1)}\right|\right\},$$

where u_1' is such that $x = h_1'(y, u_1')$.

Consider the case where the move to y is rejected. We need to find an acceptance probability $\alpha_2(x, z)$ giving detailed balance for the second-stage proposal z. As in the single-stage case, we set up a diffeomorphism between (x, u_1, u_2) and $(z, \widetilde{u_1}, \widetilde{u_2})$, where $\widetilde{u_1}$ and $\widetilde{u_2})$ would be the random numbers used in the first- and second-stage attempts from z. Then $x = h_2'(z, \widetilde{u_1}, \widetilde{u_2})$ and the first-stage move, if accepted, would have taken us to $y^\star = h_1'(z, \widetilde{u_1})$.

Completing the argument as in Section 2.2, equating integrands after making the change of variable, we find that a valid choice for the required acceptance probability is

$$\alpha_2(x, z) = \min\left\{1, \frac{\pi(z)}{\pi(x)} \frac{\widetilde{g_1}(\widetilde{u_1})\widetilde{g_2}(\widetilde{u_2})}{g_1(u_1)g_2(u_2)} \frac{[1 - \alpha_1(z, y^\star)]}{[1 - \alpha_1(x, y)]} \left|\frac{\partial(z, \widetilde{u_1}, \widetilde{u_2})}{\partial(x, u_1, u_2)}\right|\right\}. \qquad (7.1)$$

In a model-jumping problem, we would commonly take y and z to lie in the same model, and $y^\star$ to be in the same model as x, although as discussed by Green and Mira, other choices are possible. For example, where models are ordered by complexity, z might lie between x and y, so that the second-stage proposal is less 'bold'.

7.2 Efficient proposal choice for reversible jump MCMC

The most substantial recent methodological contribution to reversible jump MCMC generally is work by Brooks *et al.* (2000) on the efficient construction of proposal distributions.

This is focussed mainly on the quantitative question of selecting the proposal density ($g(u)$ in Section 2.2) well, having already fixed the transformation ($x' = h(x, u)$) into the new space. The qualitative choice of such a transformation h is perhaps more elusive and challenging.

Brooks, Giudici and Roberts propose several new methods, falling into two main classes. The first is concerned with analysis of the acceptance rate (2.3) as a function of u for small u (on an appropriate scale of measurement). The second class of methods work in a product-space formulation somewhat like that in Section 3, including some novel formulations with autoregressively constructed auxiliary variables.

Their methods are implemented and compared on examples including choice of autoregressive models, graphical gaussian models, and mixture models.

7.3 Diagnostics for reversible jump MCMC

Monitoring of MCMC convergence on the basis of empirical statistics of the sample path is important, while not of course a substitute for a good theoretical understanding of the chain. There has been some concern that across-model chains are intrinsically more difficult to monitor, perhaps implying their use should be avoided.

In truth, the degree of confidence that convergence has been achieved provided by 'passing' a diagnostic convergence test declines very rapidly as the dimension of the state space increases. In more than, say, a dozen dimensions, it is difficult to believe that a few, even well-chosen, scalar statistics give an adequate picture of convergence of the multivariate distribution. It is high, rather than variable, dimensions that are the problem.

In most trans-dimensional problems in Bayesian MCMC it is easy to find scalar statistics that retain their definition and interpretation across models, typically those based on fitted and predicted values of observations, and these are natural candidates for diagnostics, requiring no special attention to the variable dimension.

However, recognising that there is often empirical evidence that a trans-dimensional simulation stabilises more quickly within models than it does across models, there has been recent work on diagnostic methods that address the trans-dimensional problem more specifically. A promising approach by Brooks and Giudici (2000) is based on analysis of sums of squared variation in sample paths from multiple runs of a sampler. This is decomposed into terms attributable to between- and within-run, and between- and within-model variation.

Acknowledgements

Much of the work discussed here benefited strongly through HSSS funding, both of workshops and conferences where I had a chance to discuss it, and of individual visits to pursue research collaborations. I am grateful to the ESF, and to all my collaborators on reversible jump MCMC research problems: Carmen Fernández, Paolo Giudici, Miles Harkness, Matthew Hodgson, Antonietta Mira, Agostino Nobile, Marco Pievatolo, Sylvia Richardson, Luisa Scaccia and Claudia Tarantola.

References

Besag, J. (1994). Contribution to the discussion of paper by Grenander and Miller. *Journal of the Royal Statistical Society*, B, **56**, 591–2.

Besag, J. (1997). Contribution to the discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society*, B, **59**, 774.

Besag, J. (2000). *Markov chain Monte Carlo for statistical inference.* Working paper, No. 9. Center for Statistics and the Social Sciences, University of Washington.

Besag, J., Green, P. J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

Brooks, S. P. and Giudici, P. (2000). MCMC convergence assessment via two-way ANOVA. *Journal of Computational and Graphical Statistics*, **9**, 266–85.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2000). *Efficient construction of reversible jump MCMC proposal distributions.* Manuscript.

Cappé, O., Robert, C. P. and Rydén, T. (2001). *Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers.*

Preprint no. 2001:27, Centre for Mathematical Sciences, Lund University.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society*, B, **57**, 473–84.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* Chapman and Hall, London.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–21.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, **96**, 270–81.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Duane, S., Kennedy, A. D., Pendleton, B, J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physical Letters*, B, **195**, 216–22.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, B, **56**, 501–14.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–73.

Godsill, S. J. (2001). On the relationship between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics*, **10**, 230–48.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–32.

Green, P. J. and Han, X.-L. (1992). Metropolis methods, Gaussian proposals and antithetic variables. In *Stochastic models, statistical methods and algorithms in image analysis,* Lecture Notes in Statistics, No. 74, (ed. A. Frigessi, P. Barone, and M. Piccioni), pp.142–64. Springer-Verlag, Berlin.

Green, P. J. and Mira, A. (2001). Delayed rejection in reversible jump Metropolis–Hastings. *Biometrika*, **88**, 1035–53.

Green, P. J. and O'Hagan, A. (1998). *Model choice with MCMC on product spaces without using pseudo-priors.* Department of Mathematics, University of Nottingham.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society*, B, **56**, 549–603.

Han, C. and Carlin, B. P. (2001). MCMC methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, **96**, 1122–32.

Hodgson, M. E. A. (1999). A Bayesian restoration of an ion channel signal. *Journal of the Royal Statistical Society*, B, **61**, 95–114.

Hurn, M. A., Justel, A., and Robert, C. P. (2001). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics.* To appear.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, **19**, 451–8.

Neal, R. M. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". `http://www.cs.utoronto.ca/~radford`.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society*, B, **56**, 3–48.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, **60**, 607–12.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Practical Markov chain Monte Carlo*, (ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), pp.215–239. Chapman and Hall, London.

Preston, C. J. (1977). Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, **46**, 371–91.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*, B, **59**, 731–92.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society*, B, **39**, 172–212.

Roeder, K. and Wasserman, L. (1997). Contribution to the discussion of paper by Richardson and Green. *Journal of the Royal Statistical Society*, B, **59**, 782.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Annals of Applied Probability*, **8**, 1–9.

Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, **18**, 2507–15.

Waagepetersen, R. and Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward QTL-mapping. *International Statistical Review*, **69**, 49–61.

Proposal densities, and product space methods

Simon J. Godsill
University of Cambridge, UK

1 Introduction

In this article, Peter Green has provided the most informative and complete survey currently available of the issues surrounding Bayesian model uncertainty using MCMC methods. Naturally he has focussed on the reversible jump methods which have dominated the field over recent years, although he has pointed out the close relationships with the product space formulations of Besag (1997), Carlin and Chib (1995), Godsill (2001) and Dellaportas *et al.* (2002).

Practitioners have readily adopted reversible jump methods for use in complex Bayesian problems, and yet even after several years in the literature the methods have a reputation for being somehow 'difficult' to understand and still more difficult to implement successfully. Green's article helps further to demystify the reversible jump methodology by providing some useful new discussion material and a very transparent derivation of the basic results. The article also discusses recent developments in proposal design and introduces a novel proposal mechanism for general models.

So, is there any methodological work still to be done in the field? Green's article is very clear on this issue: the basic frameworks, whether pure reversible jump or combined with product space ideas, are well established; however, the specifics of a generic implementation are not, and it is clear that it is these areas that can most benefit from renewed research effort. In fact, given the general interest from a wide variety of disciplines in this topic, there have been surprisingly few methodological developments in the area up to now. In the following sections, I will focus on just two developing topics: automatic proposal generation and product space methods.

2 Construction of proposal densities

Key to the effective operation of reversible jump methods is the choice of proposal distributions. Most applications to date have constructed proposals on an ad hoc basis, attempting to place proposed parameters in regions of high probability mass in the new model's parameter space. This can be successful in some cases, but it is tempting to seek an automatic procedure that does not require the tuning and pilot runs often required in these ad hoc settings. There have been some recent advances in this direction, as discussed in Green's article. I will attempt to interpret Green's new proposal mechanism in the light of more standard Gaussian approximation methods for reversible jump.

In Godsill (2001) it was suggested that an optimal choice of proposal would be the full conditional posterior probability for the parameters in the new model, i.e. set $q(\theta_{k'}) = p(\theta_{k'}|k', y)$, in which case the acceptance ratio simplifies to

$$\frac{p(k'|y)q(k|k')}{p(k|y)q(k'|k)}$$

where $q(k'|k)$ is the probability that model k' is proposed from model k. We note that this highly idealised setting leads to an acceptance probability which is constant for all values of $\theta_{k'}$. This is in agreement with the objectives of the 'higher order methods' proposed in Brooks *et al.* (2000), in which proposals are specifically designed so that one or more derivatives of the acceptance ratio are set to zero locally at a chosen representative 'centering' point. Brooks *et al.* (2000) lend some theoretical weight to the suggestion that $p(\theta_{k'}|k', y)$ is a good proposal density by proving that the *capacitance* of the Markov chain is optimised by this choice of proposal in a simple two-model setting, and I would conjecture that the result is also valid in much more general model selection settings. This suggestion leads to the much-used idea that the proposal distribution, while in practice never *equal* to $p(\theta_{k'}|k', y)$, should be designed to approximate the full conditional if possible. A natural starting point here is a Gaussian proposal matching the 1st and 2nd order moments of the target conditional distribution. Using the same notation as Green's article, we propose a $n_{k'}$-dimensional vector v from the standard normal density, and generate the proposed parameter as $\theta_{k'} = \mu_{k'} + B_{k'}v$, giving acceptance ratio

$$A = \frac{p(k', \theta_{k'}|y)q(k|k')q(v')|B'_k|}{p(k, \theta_k|y)q(k'|k)q(v)|B_k|}. \tag{2.1}$$

In the case that the target parameter conditional is indeed Gaussian with moments $\mu_{k'}$ and $B_{k'}B_{k'}^T$ this simplifies to

$$A = \frac{p(k'|y)q(k|k')}{p(k|y)q(k|k')},$$

and we have perfectly adapted Metropolis–Hastings on the marginal model index space. Thus, in the case of a Gaussian target with correctly specified Gaussian proposals, the acceptance probabilities of this and Green's proposed method are identical and hence the two samplers explore the model indexing space equally rapidly. The interesting possibilities with Green's proposal arise when the targets are non-Gaussian, since the acceptance ratio of Green's method then appears to eliminate some of the variability in the acceptance ratio by replacing $q(v')/q(v)$ in (2.1) with a single term $q(u)$, which is the density of a generally much lower dimensional Gaussian than either $q(v)$ or $q(v')$. The question then arises as to how the target ratios $p(k', \theta_{k'}|y)/p(k, \theta_k|y)$ compare between the two approaches, and it is clear that when the target is strongly non-Gaussian, either method may well lead to high acceptance probabilities. However, this is qualitative thinking and it would be very interesting to discover how these two related approaches fared relative to one another in the examples of Section 6.1 of Peter Green's article. Clearly the approximation of each candidate model, even with a Gaussian, will require a great deal of work for large model spaces. However, one can envisage hybrid approaches in which a substantial proportion of the parameters remain fixed in model jumping proposals, as in many standard reversible jump implementations to date, while a Gaussian approximation is applied to a more manageable subset of parameters conditional on those fixed parameters.

3 Product space methods

Product space methods provide another interesting viewpoint on model uncertainty, since they allow simulation to be performed, at least conceptually, on a fixed dimension space. Various authors have shown that reversible jump algorithms can be obtained as special cases of product space methods (and vice versa); see Besag (1997), Godsill (2001) and Dellaportas *et al.* (2002).

Very general classes of model space sampling can be written in the composite model space framework of Godsill (2001), which is a product space representation, allowing for any overlap between parameters of different models that is computationally convenient (for example, nested models and variable selection models are easily encoded within the framework). Consider a 'pool' of N parameters $\theta = (\theta_1, \ldots, \theta_N)$. A candidate model k can be described in terms of this pool of parameters by means of an indexing set $\mathcal{I}(k) = \{i_1(k), i_2(k), \ldots, i_{l(k)}(k)\}$ which contains $l(k)$ distinct integer values between 1 and N. The parameters $\theta_{\mathcal{I}(k)}$ of model k are then defined as $\theta_{\mathcal{I}(k)} = (\theta_i;\, i \in \mathcal{I}(k))$. In the simplest case we have $\mathcal{I}(k) = k$, which leads to a straightforward model selection scenario with no overlap between model parameters. In other cases, such as variable selection or nested models, it may be convenient to 'share' parameters between more than one model. The posterior distribution for the composite model space can now be expressed as

$$p(k, \theta|y) = \frac{p(y|k, \theta_{\mathcal{I}(k)})\, p(\theta_{\mathcal{I}(k)}|k)\, p(\theta_{-\mathcal{I}(k)}|\theta_{\mathcal{I}(k)}, k)\, p(k)}{p(y)}, \tag{3.1}$$

where $\theta_{-\mathcal{I}(k)} = (\theta_i;\, i \in \{1, \ldots, N\} - \mathcal{I}(k))$ 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. This fixed dimensionality distribution can now be used as the target for an MCMC algorithm. One of the possible benefits of such a scheme, as suggested in Godsill (2001), is that parameters from models other than the current model can in principle be stored and used for construction of effective proposals when those other models are proposed again. There are, however, some basic pitfalls which can beset this type of approach. The first is storage: one wouldn't wish to store all parameters of all models in memory if the pool of parameters θ is large. The second is tractability. Consider the pure model selection scenario in which there is no overlap between parameters, i.e. $\mathcal{I}(k) = k$. Now, it might seem sensible to set the target density for some or all of the unused parameters equal to the data conditional posterior, in which case they can be updated at each iteration according to any suitable MCMC scheme and they will always be generating useful values for future model jumping proposals. This can be achieved by choosing the arbitrary prior distribution for these parameters as follows:

$$p(\theta_{-k}|\theta_k, k) = \prod_{j \neq k} p(\theta_j|j, y).$$

However, it is easily verified that model jumping proposals under such a scheme require the marginal model probabilities in the acceptance ratio, and hence the method is self-destroying as it requires us to know exactly one of the quantities we wish to estimate! Clearly the arbitrary prior probability should not be chosen in this intuitively reasonable way.

Another approach which might have similar benefits would be to assign some reasonable distributions for the arbitrary priors, such as a tractable approximation to the data conditional posterior distribution for those parameters, but to apply a very slowly mixing Markov chain when updating these parameters. This would allow the parameters of each model to retain some memory of their earlier configuration when that model was last selected by the MCMC. A promising approach related to this concept has been devised by Brooks *et al.* (2000). In it they assume a nested structure to the models, and augment the parameter space with sufficient auxiliary variables to make the total parameter space equal in dimensionality to the most complex candidate model. These auxiliary variables are then slowly updated at each iteration according to an autoregressive Markov chain with a standard Gaussian stationary distribution. The auxiliary variables are then used directly to generate deterministic model jumping proposals to higher order models. The extra memory and persistence introduced into the chain in this way is shown to induce a better exploration of the tails of the model order distribution for a graphical models example.

More general schemes with this flavour can easily be devised based on the general product space framework. It may be reasonable, for example, to use one or more of the auxiliary parameters to help construct a random proposal rather than a deterministic one. Another extension would address the memory storage problems: rather than update all of the auxiliary parameters using a slowly mixing Markov chain, update only those parameters within some suitably chosen 'neighbourhood' of the currently selected model. The remaining auxiliary parameters are sampled independently directly from their target distribution, which would be carefully chosen for tractability, and hence do not need to be sampled until their corresponding model number is proposed.

It seems reasonable that ideas of this sort can lead to improved performance of reversible jump algorithms. There will usually however be an increased burden of computational load and memory storage requirements, so it must remain to be seen whether performance improvements are sufficient to merit the extra work.

Trans-dimensional Bayesian nonparametrics with spatial point processes

Juha Heikkinen
Finnish Forest Research Institute, Helsinki, Finland

1 Introduction

Point processes are a class of models where the notion of variable dimension is inherent. The main part of this discussion is concerned with the application of marked point processes as prior models in nonparametric Bayesian function estimation, reformulating and revising earlier joint work with Elja Arjas and listing some other related work (Section 2). Accordingly, the discussion is centered on trans-dimensional *modelling* rather than on the simulation techniques themselves, and connects to some of the material in the chapters by Sylvia Richardson and Hurn, Husby and Rue. I shall end, however, with an example illustrating the role of the dimension-matching requirement (Section 3). The point made there is rather marginal to Green's main message, but hopefully interesting and/or instructive to modellers working with constraints.

2 Nonparametric Bayesian function estimation

Heikkinen and Arjas (1998) introduced a (trans-dimensional) nonparametric Bayesian approach to the estimation of the intensity function of a spatial Poisson process. The approach is similar to that in the change point and image analysis examples of Green (1995), and can be directly generalised to a wide variety of function estimation problems (Heikkinen 1998). It has been applied to a problem involving simultaneous interpolation, regression, and intensity estimation (Heikkinen and Arjas 1999), and closely related methods have been developed for image analysis (Nicholls 1998; Møller and Skare 2001), multivariate regression and classification (Denison *et al.* 2002b), and disease mapping (Knorr-Held and Raßer 2000; Denison and Holmes 2001). The following paragraphs show how I would now prefer to introduce the method.

Consider the estimation of real valued surfaces $f : S \to \mathcal{R}$ defined on a bounded support $S \subset \mathcal{R}^2$. A trans-dimensional approximation of f is obtained through its parametrisation by marked point pattern $\theta = \{(x_1, y_1), \ldots, (x_k, y_k)\}$, in which the locations are a simple point pattern $x = \{x_1, \ldots, x_k\}$ on S with a variable number k of randomly located points, and the marks represent values $y_i = f_\theta(x_i)$ of the approximating function. To complete the approximation, we apply some rough and simple inter/extrapolation rule to determine $f_\theta(s)$, $s \in$

$S \setminus x$. By pointwise averaging over a large number of such rough approximations, varying the number and locations of the points in x, we can then obtain a smooth estimate of f. With unbounded k the parameter space is effectively infinite-dimensional and hence the inference honestly nonparametric, yet the computations can be handled by trans-dimensional MCMC.

The inter/extrapolations of Heikkinen and Arjas (1998) were step functions on the Voronoi tessellations of S generated by the location patterns x (see Hurn *et al.* this volume, Section 2.2; S. Richardson, this volume, Section 3.1). However, Voronoi tessellations could be replaced by the Delaunay or other more general triangulations (cf. Nicholls 1998). In addition to the more flexible geometry, triangular partitions offer the opportunity of making the function approximations piecewise linear instead of piecewise constant (see below). In the estimation of smooth functions, I would prefer the computationally simpler Delaunay triangulations, the greater flexibility of other triangulations being more valuable in problems like segmentation (Nicholls 1998).

Our prior of x was the homogeneous Poisson process, and large differences between nearby function values were penalised by a Markov random field prior for $y|x$. This led to unnecessary complications with the normalising constants, which could have been avoided by modelling the marked point pattern θ directly as a nearest-neighbour Markov point process with correlated marks, as did Møller and Skare (2001). Then the marginal prior of x is no longer a Poisson process, but that seems like a small price to pay for an otherwise more tractable model. Although Denison and Holmes (2001) deem this smoothing unnecessary in the first place, I think that it should lead to qualitatively more reasonable individual approximations of f, and thereby to more realistic inferences on its shape, for example. For extrapolations beyond the convex hull of the data, dependence priors seem essential.

Motivated by such considerations, let me then sketch an approach I would currently suggest. Assuming S to be a polygon, let θ be a marked point process including locations on the edges and vertices of S as in the model of Nicholls (1998). Define the prior density of θ with respect to the distribution of the appropriate marked Poisson process by something like

$$p(\theta) \propto \lambda^k \exp\left\{-\tau \sum_{x_i \sim_x x_j} (y_i - y_j)^2\right\},$$

where $x_i \sim_x x_j$, if the tiles $S(x; x_i) \ni x_i$ and $S(x; x_j) \ni x_j$ of the Voronoi tessellation generated by x are adjacent. Finally, define f_θ as that unique surface which passes through all points (x_i, y_i) of θ and is linear within each triangle in the Delaunay tessellation generated by x. If f can only take a finite (and small) number of distinct values, as in image classification, for example, I would follow Møller and Skare (2001) in using the Voronoi step functions and an extension like

$$p(\theta) \propto \lambda^k \exp\left\{-\tau \sum_{x_i \sim_x x_j} \mathbf{1}(y_i \neq y_j)\right\}$$

of the Potts model, where **1** denotes the indicator function.

Theoretically, this approach works regardless of the dimension of S. However, the effort needed both for the implementation and for the computations increase rapidly with the dimension; for an example, see the 3-d problem in reservoir modelling tackled by Møller and Skare (2001). Independence priors allow for a computationally feasible approach for moderate dimensional S (Denison *et al.* 2002b), but Denison *et al.* (2002a) have found that they do not work well in very high dimensions, either.

3 On constraints and dimension matching

In most applications of trans-dimensional MCMC, the major problem seems to be finding *efficient* proposal distributions. When there are constraints in the parameter space, however, even the choice of *valid* proposals may not be trivial. A typical case in the function estimation context are problems involving interpolation (Heikkinen and Arjas 1999), of which a toy example is given below.

Consider the function estimation problem of Section 2 with the constraint $f(s_0) = f_0$ for some $s_0 \in S$, and step function approximations f_θ taking constant value y_i on each Voronoi tile:

$$f_\theta(s) = \sum_{i=1}^{k} y_i \mathbf{1}\{s \in S(x; x_i)\}.$$

Suppose we wish to implement the simplest possible sampler with two kinds of move proposal: death of one random point in the current θ and birth of a new point (ξ, η) with a uniform random location $\xi \in S$ and mark η sampled from some distribution on $\mathcal{R}$. In Green's formalism, we would then have $r = 2k$, $r' = 2k + 2$, $u = (\xi, \eta)$, and $d' = 0$ for the birth move from θ with k points to $\theta' = \theta \cup \{u\}$. But if $s_0 \in S(x \cup \{\xi\}; \xi)$, then the only proposal yielding a positive acceptance probability would be the (one-dimensional) $\theta' = \theta \cup (\xi, f_0)$. This would leave the mark η of u unused and hence violate the dimension-matching requirement.

The concrete consequences of the failure in dimension-matching are revealed only when trying to work out the acceptance probability for the death of (x_i, y_i) with $s_0 \in S(x; x_i)$. For positive chances of acceptance, we are forced to propose function value f_0 on that tile $S(x \setminus x_i; x_j)$ which contains s_0 in the proposed tessellation. In other words, the death proposal is

$$\theta' = \theta \setminus \{(x_i, y_i), (x_j, y_j)\} \cup (x_j, y_j'),$$

where $y_j' = f_0$. But if $y_j \neq f_0$, then our simple sampler cannot reverse this move, because it proposes $y_j'' = y_j' = f_0$ in the birth move from θ'.

Returning to the birth move $x' = x \cup \xi$ with $s_0 \in S(x'; \xi)$, the considerations above lead to the conclusion that the dimension we cannot use in proposing the function value on tile $S(x'; \xi)$, that is, the random mark η of u, must be used to perturb the current function value $y_j (= f_0)$ on the tile $S(x; x_j)$ containing s_0. See Heikkinen and Arjas (1999) for the details of one such sampler.

One might ask, why not just keep a fixed generating point (s_0, f_0) and avoid the whole difficulty, but this would result in different smoothing around s_0 than elsewhere.

Additional references in discussion

Denison, D. G. T. and Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, **57**, 143–9.

Denison, D. G. T., Adams, N. M., Holmes, C. C., and Hand, D. J. (2002a). Bayesian partition modelling. *Computational Statistics and Data Analysis*, **38**, 475–85.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002b). *Bayesian methods for nonlinear classification and regression*, Chapter 7. Wiley, Chichester.

Heikkinen, J. (1998). Curve and surface estimation using dynamic step functions. In *Practical nonparametric and semiparametric Bayesian statistics*, (ed. D. Dey, P. Müller, and D. Sinha), pp.255–72. Springer-Verlag, New York.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**, 435–50.

Heikkinen, J. and Arjas, E. (1999). Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, **55**, 738–45.

Holmes, C. C., Denison, D. G. T., and Mallick, B. K. (1999). *Bayesian partitioning for classification and regression.* Technical report, Department of Mathematics, Imperial College, London.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**, 13–21.

Møller, J. and Skare, Ø. (2001). Coloured Voronoi tessellations for Bayesian image analysis and reservoir modelling. *Statistical Modelling*, **1**, 213–32.

Nicholls, G. (1998). Bayesian image analysis with Markov chain Monte Carlo and colored continuum triangulation models. *Journal of the Royal Statistical Society*, B, **60**, 643–59.