Trans-dimensional Markov chain Monte Carlo

Peter J. Green[*]
University of Bristol, UK.

Partial draft – 8 November 2000

(to be discussed by Simon Godsill and Juha Heikkinen)

Summary

In the context of sample-based computation of Bayesian posterior distributions in complex stochastic systems, this chapter discusses some of the uses for a Markov chain with a prescribed invariant distribution whose support is a union of euclidean spaces of differing dimensions. This leads into a re-formulation of the reversible jump MCMC framework for constructing such 'trans-dimensional' Markov chains. This framework is compared to alternative approaches for the same task, including methods that involve separate sampling within different fixed-dimension models. We consider some of the difficulties researchers have encountered with obtaining adequate performance with some of these methods, attributing some of these to misunderstandings, and offer tentative recommendations about algorithm choice for various classes of problem. The chapter concludes with a look towards desirable future developments.

Some key words: Bayes factors, Bayesian model selection, Delayed rejection, Jump diffusion, Metropolis-Hastings algorithm, Reversible jump methods, Simulated tempering.

1 Introduction

For the audiences of this book, or the workshop that produced it, it should be unnecessary to assert the huge 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, ...) 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 I will call '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 abstractly as a matter of joint inference about a model indicator k and a parameter vector θ_k, where the

[*] Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: P.J.Green@bristol.ac.uk.

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. But, how to compute it?

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.

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 a 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, often arising in Bayesian nonparametrics, k is unlikely to be of direct inferential interest.

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-Hasting 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 whole point of the 'reversible jump' framework is to render the measure theory completely invisible, by means of a construction using only ordinary densities. In fact, in the formulation given below, improving 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, and will require its transition kernel P to satisfy the detailed balance condition, written in integral form as

$$\int_{(x,x')\in A\times B} \pi(\mathrm{d}x)P(x,\mathrm{d}x') = \int_{(x,x')\in A\times B} \pi(\mathrm{d}x')P(x',\mathrm{d}x) \tag{1}$$

for all Borel sets $A, B \subset \mathcal{X}$. In Metropolis-Hastings, we make a transition by first drawing a proposed 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 makes an equal contribution to each side of equation (1), so can be neglected, and we are left with the requirement

$$\int_{(x,x')\in A\times B} \pi(\mathrm{d}x)q(x,\mathrm{d}x')\alpha(x,x') = \int_{(x,x')\in A\times B} \pi(\mathrm{d}x')q(x',\mathrm{d}x)\alpha(x',x). \tag{2}$$

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 $\mathcal{X} \times \mathcal{X}$, and has density (Radon-Nikodym derivative) f with respect to this μ. Then (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 appropriate A, B if

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

If we wrote this rather more informally as

$$\alpha(x,x') = \min\left\{1, \frac{\pi(\mathrm{d}x')q(x',\mathrm{d}x)}{\pi(\mathrm{d}x)q(x,\mathrm{d}x')}\right\} \tag{3}$$

then the similarity with the usual expression using densities is apparent.

2.2 An explicit representation in terms of random numbers

This may seem rather abstract. Fortunately, in most cases, the dominating measure and Radon-Nikodym derivatives can be generated almost automatically, by considering how the transition will actually be implemented. 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 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. We can then re-express the left-hand side of (2) as an integral with respect to (x, u), and it becomes:

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

Now consider how 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 re-express the right-hand side of (2) as an integral with respect to (x', u'), and then apply the standard change-of-variable formula, to see that the $(d + r)$–dimensional integral equality (2) holds if

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

whence a valid choice for α is

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')g'(u')}{\pi(x)g(u)}\left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\}. \tag{4}$$

It is often easier to work with this expression than the usual one.

This reversible jump formalism perhaps appears a little indirect, but 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, where this subset need not be parallel to coordinate axes. Thus the proposal measure can be singular. Note that there is a deliberate redundancy in separating the generation of the random innovation u and the calculation of the proposal value through the deterministic function $x' = h(x, u)$. This allows the same proposal distribution to be expressed in many different ways, a choice that can be exploited to the user's convenience.

2.3 The trans-dimensional case

The real bonus of this formalism, however, is that expression (4) applies, without change, in a variable dimension context. Provided that the transformation from (x, u) to (x', u') remains a diffeomorphism, the individual dimensions of x and x' (and hence of u and u') 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'.

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

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) 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 (1) is then satisfied.

The analysis leading to (3) and (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 equilibrium joint state/proposal distribution, the *equilibrium joint state/proposal distributions.* That the degrees of freedom in this joint distribution 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, Green, Higdon and Mengersen (1995, appendix ?)). Similarly, the Jacobian comes into the acceptance probability simply through the fact that the proposal destination $x' = h(x, u)$ is specified indirectly, a formalism that gives singular moves a natural expression; it has nothing to do with the jump in dimension.

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 it 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 is of course exactly equivalent.

Other authors have provided different pedagogical descriptions of reversible jump. Waagepetersen and Sorensen (2000) 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

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

Jump diffusion. In the context of a challenging computer vision application, Grenander and Miller (1994) proposed a sampling strategy they termed jump diffusion. This was composed of 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 in fact corrected for the time discretisation by a Metropolis-Hastings accept/reject decision (giving a so-called Metropolis-adjusted Langevin algorithm or MALA) (Besag, 1994; Roberts?, 199?), 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. Other authors (???), however, have found jump-diffusion more complicated to set up and adjust than reversible jump. The approach probably deserves further study, particularly to extend the very limited options for the between-model moves: Grenander and Miller, and Phillips and Smith, consider only 'Gibbs' and 'Metropolis' jump dynamics.

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). Much 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, and borrows the birth-and-death simulation idea to give a very promising methodology for finite mixture analysis. He also notes that the approach appears to have much wider application, citing change point analysis and regression variable selection as partially worked examples. The key features of these three settings that allow the approach to work seem to be the possibility of integrating out latent variables so that the likelihood is fully available, and the nested structure of the models under consideration. It is thus much more narrow in scope than reversible jump, but still surprisingly powerful.

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

$$p(k)p(\theta_k|k)p(\theta_{-k}|k, \theta_k)p(Y|k, \theta_k) \tag{5}$$

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: $p(\theta_{-k}|k, \theta_k) = \prod_{l\neq k} p(\theta_l|k)$ (with $p(\theta_l|k)$ not depending on k for $k \neq l$), and 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, Forster and Ntzoufras (2000) 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 (2000) provides a general 'composite model space' framework that embraces all of these methods, including reversible jump, facilitating comparisons between them. He devised the formulation (5), 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 provides 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

Within-model simulation. Here we conduct simulations within each model k separately. The posterior for the parameters θ_k is in any case a within-model notion:

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

and is the target for an ordinary Bayesian MCMC calculation for model k.

As for the posterior model probabilities, 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 term being the *Bayes factor* for model k_1 vs. k_0), 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.

Estimating the marginal likelihood. There are various possible estimates based on importance sampling, some of which are well-studied, for example

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

based on a MCMC sample $\theta_k^{(1)}, \theta_k^{(2)}, \ldots$ from the posterior $p(\theta_k|Y,k)$, or

$$\widehat{p}_2(Y|k) = N^{-1}\sum_{t=1}^{N} p(Y|k,\theta_k^{(t)})$$

based on a sample from the *prior* $p(\theta_k|k)$. Both of these has its faults, and composite estimates can perform better. See, for example, Newton and Raftery (1994) and Gelfand and Dey (1994).

- reviews: Carlin and Louis (2nd ed, sec 6.3.1), Han and Carlin
- Ritter and Tanner?, Chib(1995), Chib and Jeliazkov (2000)
- Meng and Wong, Mira and Nicholls

5 Comparative discussion of scope, limitations and difficulties

Discussion of strengths and weaknesses of RJMCMC and the other trans-dimensional setups above compared to within-model simulations that compute marginal likelihoods and thence Bayes factors.

- Han and Carlin (although this really has the very limited objective of computing Bayes factors)
- RJ 'difficult to tune': qualitative arguments/choices.
- how many models and how homogeneous are they – a major factor in usefulness of model-jumping MCMC

6 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 Figure 9 (reproduced as Figure 1 below), 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 (2000) 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' model are completely wasted from the point of view of extracting useful posterior information; let us try to analyse when it will nevertheless be worthwhile to use a trans-dimensional sampler.

6.1 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 $\pi_2 D_2 + \pi_1 Q_{12} = \pi_2$.

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, assuming that we make a random choice between the two moves available from each state: α and β are the probabilities of choosing to attempt the between-model move in models 1, 2 respectively. The invariant distribution is easily seen to be $\pi = (\gamma\pi_1, (1-\gamma)\pi_2)$, where γ satisfies the equation $\gamma\alpha = (1-\gamma)\beta$.

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.

6.2 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 a 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, so that $(Q_{11})_{i,i\pm1} = 0.03$, $(Q_{12})_{i,i\pm1} = 0.2 \times 0.1$, $(Q_{21})_{i,i\pm1} = 0.2 \times 0.1$, 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.

6.3 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. 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 prior on 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.

Another example is illustrated in Figure 1.

7 Generic strategies for constructing jump proposals

Given that reversible jump is merely a formulation of quite general Metropolis-Hastings for state spaces like $\mathcal{X} = \bigcup_{k\in\mathcal{K}}(\{k\} \times \mathcal{R}^{n_k})$, and that Metropolis-Hastings is surely a very promising and

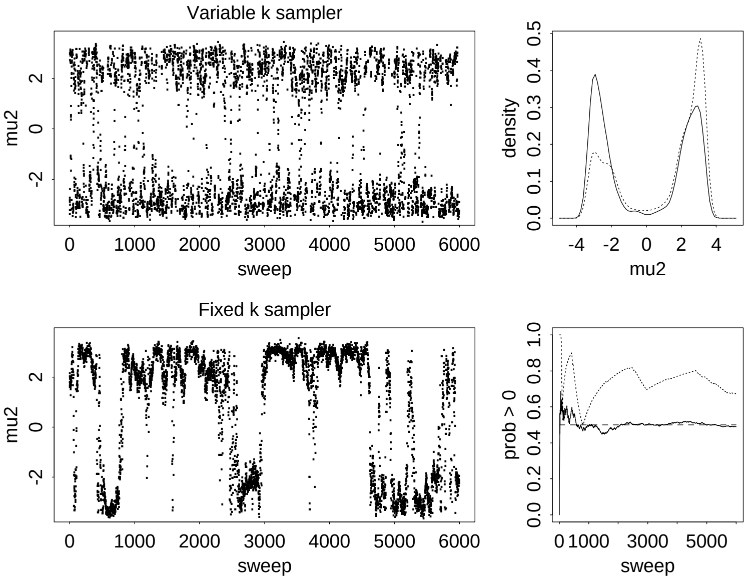

Figure 1: Example of tempering-by-embedding (from Richardson and Green, 1997): comparison of fixed-k and variable-k samplers for a normal mixture problem with k components, applied to the model with $k = 3$.

general paradigm for building Markov chains with given invariant distributions, one has to ask: why have some researchers found it so difficult to set up reversible jump methods for complex problems?

The answer must be to do with the particular choices of move that are being attempted. Most current methods have a split/merge or birth/death structure, and perhaps such choices are just too restrictive. Split/merge and birth/death are used for a variety of reasons: there are some successful examples to use as models, they are natural in many problems and fairly easy to think about, and they often exploit factorisations of the joint probabilities to give computationally cheap proposals.

However, let us set such considerations aside, and attempt the construction of generic moves that have a reasonably high chance of success, irrespective of cost.

7.1 Two linear models

We will take a very stylised situation, hoping to be able to visualise the structure of the problem. Consider two candidate linear regression models for an n-vector response variable y. Each includes also an offset, but we assume variances fixed at 1. The models are written as

$$y \sim N(x_0 + X\beta, I_n) \qquad \text{and} \qquad y \sim N(z_0 + Z\gamma, I_n)$$

respectively, and we take diffuse priors on the regression parameters, and equal prior probabilities on the two models. We call these the X and Z models, instead of model 1 and model 2, and we suppose both the matrices X and Z are of full rank, p and q respectively.

A state in, say, the X model may be interchangeably described by either the parameter vector β or the 'fitted value' vector $x_0 + X\beta$: we use the latter. From this point, to what point in the Z model should we propose to jump? A first thought is this should be a random perturbation to the *nearest* point in the Z model, that is, the projection of $x_0 + X\beta$ onto the hyperplane $z_0 + Z\gamma$, namely $z_0 + P_Z(x_0 + X\beta - z_0)$, where P_Z is the usual orthogonal projection operator $Z(Z^TZ)^{-1}Z^T$. See Figure 2 (left). Indeed, when the model hyperplanes are approximately parallel, this may be quite successful. Such a move can be viewed as analogous to the split/merge of steps used, for example, in the first (change point) example in Green (1995). However, in general, such a strategy ignores a crucial factor, namely whether or not $x_0 + X\beta$ is a good fit to the data y: it takes no account of the residual $y - x_0 - X\beta$. It also has the unattractive and probably undesirable property that, ignoring the random perturbations, two successive jump moves do not lead back to the same state.

Our preferred strategy is defined as follows. Let $\widehat{y}_X$ denote the best-fitting X model, in a least-squares sense: $\widehat{y}_X = x_0 + P_X(y - x_0)$, and similarly let $\widehat{y}_Z = z_0 + P_Z(y - z_0)$. Our intuition is that from $\widehat{y}_X$, a good point to jump to in the Z model would be $\widehat{y}_Z$. To obtain a proposal distribution centred at this point, we use the innovation variables u to perturb our starting point orthogonally away from the X model hyperplane, before projecting it onto the Z hyperplane. For other current states, we follow natural linear invariances. Thus our proposed new state in the Z model, given a general current state $x_0 + X\beta$ in the X model is

$$z_0 + Z\gamma = \widehat{y}_Z + P_Z\{x_0 + X\beta - \widehat{y}_X + (I_n - P_X)u\}.$$

The reverse move would propose to take $z_0 + Z\gamma$ to

$$x_0 + X\beta = \widehat{y}_X + P_X\{z_0 + Z\gamma - \widehat{y}_Z + (I_n - P_Z)u'\}.$$

See Figure 2 (centre and right). We will suppose u, u' to be spherical gaussian variables in n dimensions: $u \sim N(0, \sigma_X^2 I_n)$ and $u' \sim N(0, \sigma_Z^2 I_n)$. On the face of it, this proposal mechanism fails to respect the dimension-matching constraint, but note that many components of u are not used: in fact γ depends only on $P_Z(I_n - P_X)u$. The rank of $P_Z(I_n - P_X)$ is $q - t$, where t is the dimension of the intersection of the column spaces of X and Z: $t \leq \min(p, q)$. Thus we have dimension matching (see Section 2.3) with $d = p$, $r = q - t$, $d' = q$ and $r' = p - t$.

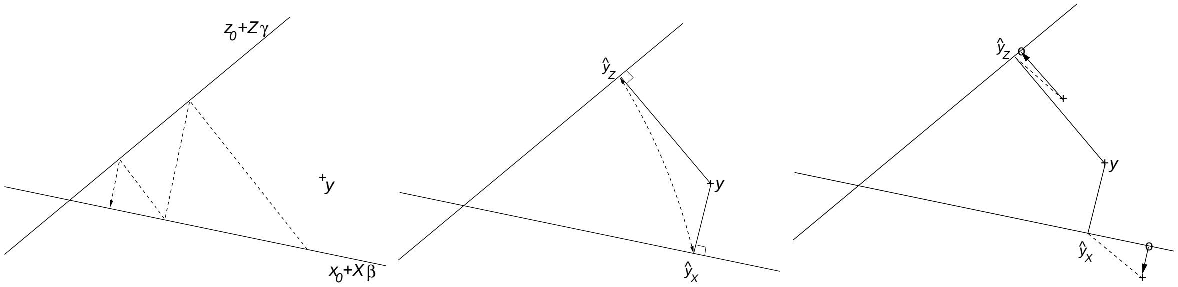

Figure 2: Illustrating model-jumping for two linear models. Left panel: proposing to jump to the nearest point in the other model; centre panel: our preferred strategy; right panel: an example, showing random perturbation away from the current model hyperplane, and projection onto the other model hyperplane.

If we define A to be a $n \times r$ matrix with $AA^T = P_Z(I_n - P_X)P_Z$ and B to be $n \times r'$ with $BB^T = P_X(I_n - P_Z)P_X$, then $P_Z(I_n - P_X)u$ can be generated as Av and $P_X(I_n - P_Z)u'$ as Bv', where $v \sim N(0, \sigma_X^2 I_r)$ and $v' \sim N(0, \sigma_Z^2 I_{r'})$. The mapping from (β, v) to (γ, v') is nonsingular, and the Jacobian of transformation is

$$\left|\frac{\partial(\gamma, v')}{\partial(\beta, v)}\right| = \left|\begin{array}{cc} (Z^TZ)^{-1}Z^TX & (Z^TZ)^{-1}Z^TA \\ (B^TB)^{-1}B^T(I_n - P_XP_Z)X & -(B^TB)^{-1}B^TP_XA \end{array}\right|$$

which, rather extraordinarily, simplifies to

$$\left(\frac{|X^TX|}{|Z^TZ|}\right)^{1/2}!$$

The acceptance probability for the proposed jump from $x_0+X\beta$ to $z_0+Z\gamma$ is therefore $\min\{1, R\}$ where

$$R = \frac{\exp(-0.5||y - z_0 - Z\gamma||^2)}{\exp(-0.5||y - x_0 - X\beta||^2)}(2\pi)^{(r-r')/2}\frac{\sigma_X^r}{\sigma_Z^{r'}}\frac{\exp(-0.5||v'||^2/\sigma_Z^2)}{\exp(-0.5||v||^2/\sigma_X^2)}\left(\frac{|X^TX|}{|Z^TZ|}\right)^{1/2}.$$

Further light on this is shed by considering some special cases and an example.

Nested models. Let us consider how this mechanism works in the case where model X is nested within model Z: that is, $x_0 = z_0$ and $X = ZS$ for some full rank $q \times p$ matrix S. Of course, $t = p < q$, so $r = q - p$ and $r' = 0$: the proposed Z to X jump is deterministic. After some manipulation, noting that $P_X P_Z = P_X$, we find that the proposal is

$$x_0 + X\beta = x_0 + P_X Z\gamma,$$

in fact, simply a projection onto the X model. In the reverse direction, since $P_Z(I_n - P_X)P_Z = AA^T$ simplifies to $P_Z - P_X$, we have

$$x_0 + Z\gamma = x_0 + X\beta + (P_Z - P_X)(y - x_0) + Av.$$

Orthogonal models. An opposite extreme arises where the models are perfectly orthogonal: $X^TZ = 0$. Then $t = 0$, and so $r = q$ and $r' = p$. The proposed moves simplify to

$$z_0 + Z\gamma = \widehat{y}_Z + Av \qquad \text{and} \qquad x_0 + X\beta = \widehat{y}_X + Bv'$$

where AA^T and BB^T are just P_Z and P_X. In this case, the method becomes an instance of Independence Metropolis-Hastings: the proposal does not depend on the current state, and this makes sense since in this orthogonal set-up, the current state provides no information to discriminate between parameters in the other model.

An example. Here is a demonstration example on a small scale. The models are for n independent observations (x_i, y_i) from

$$y_i \sim N(\beta_1 + \beta_2 x_i, 2^2) \qquad \text{or} \qquad y_i \sim N(\gamma_1 + \gamma_2(x_i - 2)^2, 2^2),$$

where x_i are evenly spaced from 0 to 4. Apart from the intercept, these two models are orthogonal. Figure 3 shows some aspects of the performance of 'nearest point' sampler and our preferred alternative; the former needed much larger proposal spreads σ and σ' for acceptable mixing, but is nevertheless inferior to the latter.

7.2 Consequences for general models

Taken literally, this setting of two linear models is too special to be of any statistical interest. However, this construction may have wider significance that it seems. Provided that the data are sufficiently informative about parameters in both models, and that standard regularity conditions apply, *any* two models are locally linear, so may be expanded about suitable $\widehat{y}_X$ and $\widehat{y}_Z$, presumably the posterior modes in each model. Although approximately linear, the geometry will change somewhat from the special case, since we need to account for general (and model-specific) posterior variance structures.

An additional difficulty is that of course in a realistic setting, posterior means or modes and variances are not available – which is why we are using MCMC!

Nevertheless, we intend to explore these issues in later work, including the use of approximations based on pilot runs, perhaps further approximated for computational efficiency.

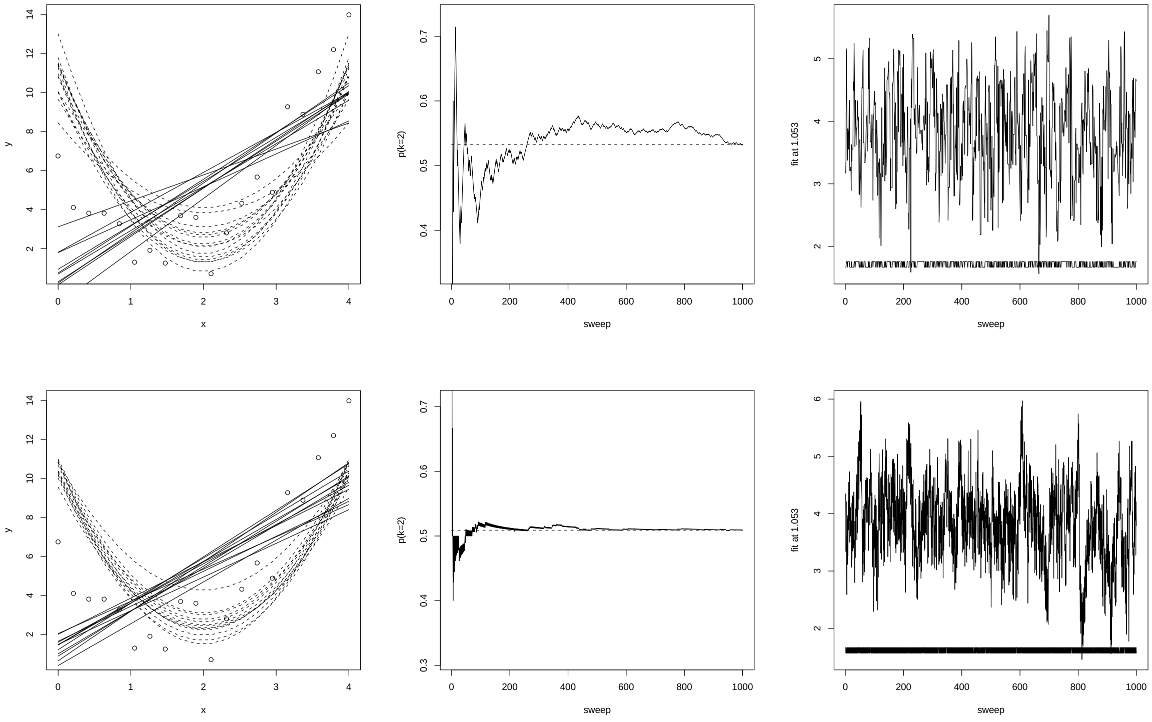

Figure 3: Comparisons between samplers for jumping between two simple regression models. Top line: jumping to the nearest point in the other model; bottom line: our preferred strategy. Left: data and samples from posterior models; centre: convergence of cumulative frequency of model Z; right: trace plot for value of regression curve at a given x.

8 Methodological extensions

- Delayed rejection
- Diagnostics for RJMCMC
- Automatic proposal choice for RJMCMC (Roberts, Brooks and Giudici)

9 Future directions

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 MCMC research problems: Robert Aykroyd, Julian Besag, Steve Brooks, Carmen Fernández, Arnoldo Frigessi, Paolo Giudici, Xiao-liang Han, Miles Harkness, David Higdon, Matthew Hodgson, Kerrie Mengersen, Antonietta Mira, Duncan Murdoch, Agostino Nobile, Marco Pievatolo, Sylvia Richardson, Claudia Tarantola, and Iain Weir.

References

Besag, J. (1994) Contribution to the discussion of paper by Grenander and Miller. *Journal of the Royal Statistical Society*, B, **56**, 000–000.

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.* Department of Statistics, University of Washington.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society*, B, **55**, 25–37.

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

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society*, B, **57**, 473–484.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2000) *Marginal likelihood from the Metropolis-Hastings output.* Manuscript.

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935–944.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2000) *On Bayesian model and variable selection using MCMC.* Department of Statistics, Athens University of Economics and Business.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, B, **56**, 501–514.

Geyer, C. J. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Godsill, S. J. (2000) On the relationship between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics*, accepted for publication.

Green, P. J. (1994) Contribution to the discussion of paper by Grenander and Miller. *Journal of the Royal Statistical Society*, B, **56**, 589–590.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Mira, A. (2000) *Delayed rejection in reversible jump Metropolis-Hastings.* Department of Mathematics, University of Bristol.

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. (2000) *MCMC methods for computing Bayes factors: a comparative review.* Division of Biostatistics, University of Minnesota.

Heikkinen, J. and Arjas, E. (1998) Nonparametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**, 435–450.

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

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.

Norman, G. E. and Filinov, V. S. (1969) Investigation of phase transitions by a Monte Carlo method. *High Temperature*, **7**, 216–222.

O'Hagan, A. (1994) *Bayesian Inference (Kendall's Advanced Theory of Statistics,* **2 B***)*, Wiley, New York.

Phillips, D. B. and Smith, A. F. M. (1996) Bayesian model comparison via jump diffusions. Chapter 13 (pp. 215–239) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Preston, C. J. (1977) Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, **46** (2), 371–391.

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**, 731–792.

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

Roberts, G. O., Brooks, S. P. and Giudici, P. (2000)

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.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (1998) *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models.* Unpublished manuscript.

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.

Waagepetersen, R. and Sorensen, D. (2000) *A tutorial on reversible jump MCMC with a view towards applications in QTL-mapping*, Department of Mathematics, Aalborg University.