Trans-dimensional Markov chain Monte Carlo

by Peter Green
(University of Bristol,
P.J.Green@bristol.ac.uk
http://www.stats.bris.ac.uk/∼peter).

(Thanks to all my collaborators on
trans-dimensional MCMC problems:
Carmen Fernández, Paolo Giudici, Miles Harkness,
David Hastie, Matthew Hodgson, Antonietta Mira,
Agostino Nobile, Marco Pievatolo,
Sylvia Richardson, Luisa Scaccia and
Claudia Tarantola.)

Trans-dimensional Markov chain Monte Carlo

What if 'the number of things you don't know is one of the things you don't know'?

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.

Formulate generically as 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.

Usually in a frequentist setting, inference about these two kinds of unknown is based on different logical principles.

There may be debate on what to do with it, but the Bayesian needs only the joint posterior $p(k, \theta_k|Y)$. How should we compute it?

Hierarchical model

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, 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 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

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

– the product of posterior model probabilities and model-specific parameter posteriors.

– very often the basis for reporting the inference, and in some of the methods mentioned below is also the basis for computation.

Note the generality of this basic formulation: it embraces both

- 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 which case, k is unlikely to be of direct inferential interest).

Compatibility across models

Some would argue that responsible adoption of this Bayesian hierarchical model presupposes that, e.g., $p(\theta_k|k)$ should be compatible in 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 (how?).

But it is philosophically tenable that no such compatibility is present, and we shall not assume it.

Non-Bayesian uses

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.

Across- and within-model simulation

Two main approaches:

- *across*: one MCMC simulation with states of the form (k, θ_k)
- *within*: separate simulations of θ_k for each k.

Across-model simulation

Reversible jump MCMC

The state space for an across-model simulation is $\{(k, \theta_k)\} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{R}^{n_k})$.

Mathematically, this is not a particularly awkward object. But at least a little non-standard!

We use Metropolis-Hastings to build a suitable reversible chain.

On the face of it, this requires measure-theoretic notation, which may be unwelcome! The point of the 'reversible jump' framework is to render the measure theory invisible, by means of a construction using only ordinary densities. Even the fact that we are jumping dimensions becomes essentially invisible!

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

$$\int \pi(\mathrm{d}x) P(x, \mathrm{d}x') = \int \pi(\mathrm{d}x') P(x', \mathrm{d}x)$$

(both integrals over $(x, x') \in A \times B$),
for all Borel sets $A, B \subset \mathcal{X}$.

Compare this with

$$\pi(x) P(x, x') = \pi(x') P(x', x)$$

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 the DB equation; subtracting this leaves

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

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

Now $\pi(\mathrm{d}x) q(x, \mathrm{d}x')$ 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 DB requires

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

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

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

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

A constructive representation in terms of random numbers

Now let's get rid of this abstraction!

Consider how the transition will be implemented; we find the dominating measure and Radon-Nikodym derivatives can be generated implicitly.

Assume $\mathcal{X} \subset \mathcal{R}^d$, and that 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 a deterministic function of the current state and the random numbers: $x' = h(x, u)$, say.

The reverse transition from x' to x would be made with the aid of random numbers $u' \sim g'$ giving $x = h'(x', u')$.

The equilibrium probability of jumping from A to B is then an integral with respect to (x, u):

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

The equilibrium probability of jumping from B to A is an integral with respect to (x', u'):

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

If the transformation from (x, u) to (x', u') is a diffeomorphism (the transformation and its inverse are differentiable), then we can apply the standard change-of-variable formula, to write this as an integral with respect to (x, u).

Detailed balance says the two integrals are equal: it 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

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

involving only ordinary joint densities.

What's the point?

Perhaps a little indirect!

– but a flexible framework for constructing quite complex moves using only elementary calculus.

The possibility that $r < d$ covers the typical case 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.)

Deliberate redundancy

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 $q(x, B) = \int_{h(x,u)\in B} g(u)\mathrm{d}u$ to be expressed in many different ways, for the convenience of the user.

The trans-dimensional case

But the main benefit of this formalism is that

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

applies, without change, in a variable dimension context.

(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 'invisible'.

Dimension matching

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.

Dimension matching is necessary but not sufficient.

Details of application to model-choice

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 becomes

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

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

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 $x \notin B$, $P(x, B) = \sum_M \int_B q_m(x, \mathrm{d}x') \alpha_m(x, x')$.

The acceptance probability derivation 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.

Toy example

..... of no statistical use at all!

Suppose x lies in $\mathcal{R} \cup \mathcal{R}^2$: $\pi(x)$ is a mixture:
with probability p_1, x is $U(0, 1)$,
with probability p_2, it is Uniform on the triangle $0 < x_2 < x_1 < 1$.

I will use three moves:

(1) within $\mathcal{R}$: $x \to U(x - \epsilon, x + \epsilon)$, suppressing moves outside $(0, 1)$.

(2) within $\mathcal{R}^2$: $(x_1, x_2) \to (1 - x_2, 1 - x_1)$.

(3) between $\mathcal{R}$ and $\mathcal{R}^2$

In $\mathcal{R}$, choose (1) or (3) with probabilities $1 - r_1, r_1$.
In $\mathcal{R}^2$, choose (2) or (3) with probabilities $1 - r_2, r_2$.
Thus $j_3(x) = r_1$ for all $x \in \mathcal{R}$ and $j_3(x') = r_2$ for all $x' \in \mathcal{R}^2$.

Dimension-changing with move (3)

Proposal:
To go from $x \in \mathcal{R}$ to $(x_1, x_2) \in \mathcal{R}^2$, draw u from $U(0, 1)$ [so $g_3(u) = 1$ if $0 < u < 1$] and propose $(x_1, x_2) = (x, u)$. For reverse move, no u' required [write $g'_3(u') \equiv 1$] and set $x = x_1$. This certainly gives a bijection: $(x, u) \leftrightarrow (x_1, x_2)$, with Jacobian $= 1$.

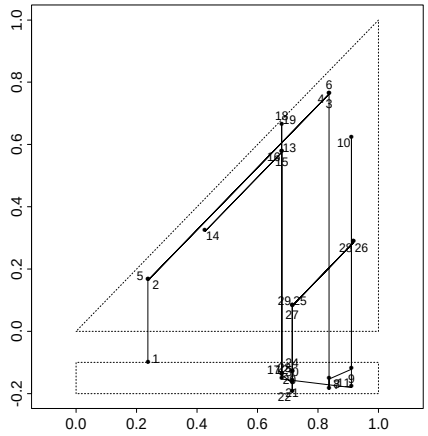
Acceptance decision:

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi(x')}{\pi(x)} \frac{j_3(x')}{j_3(x)} \frac{g'_3(u')}{g_3(u)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\} \\
&= \min\left\{1, \frac{p_2 2I[x_2 < x_1]}{p_1} \frac{r_2}{r_1} \frac{1}{1} |1|\right\} \\
&= \min\left\{1, \frac{2p_2 r_2}{p_1 r_1}\right\} I[u < x]
\end{aligned}$$

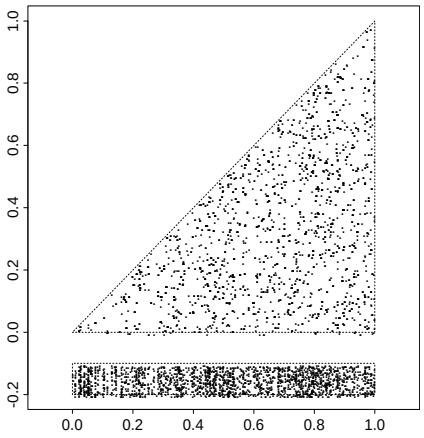
For reverse move, $\alpha = \min\{1, (p_1 r_1)/(2p_2 r_2)\}$.

First 30 steps

$p_1 = 0.4$, $p_2 = 0.6$, $r_1 = 0.7$, $r_2 = 0.4$, $\epsilon = 0.3$.



5000 steps



Some remarks and ramifications

- key role of joint state-proposal equilibrium distributions $\pi(\mathrm{d}x)q(x, \mathrm{d}x')$
- insights into Metropolis-Hastings applying quite generally
  - state-dependent mixing permissible if move probabilities enter into the acceptance probability calculation
  - contrast between this *randomised* proposal mechanism, and related *mixture* proposals
  - (contrary to some accounts that connect it with the jump in dimension) the Jacobian comes into the acceptance probability only because the proposal destination $x' = h(x, u)$ is specified indirectly

- nested models: RJ $\equiv$ proposals with atoms and usual M-H formula
- there are alternative derivations and descriptions, e.g. Waagepetersen and Sorensen (2001) and Besag (1997, 2000) (giving a novel formulation in which variable dimension notation is circumvented by augmenting x by u)
- RJ is only Metropolis-Hastings (so if it doesn't seem to work....)

Relations to other across-model approaches

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

Jump diffusion

Grenander and Miller (1994): two kinds of move – between-model jumps, and within-model diffusion using a Langevin stochastic differential equation (+ discrete-time approximation = a trans-dimensional Markov chain).

Had they corrected for the time discretisation by a M-H accept / reject decision (Metropolis-adjusted Langevin algorithm), this would have been an example of reversible jump.

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

Point processes, with and without marks

Point processes: natural example of a variable-dimension distribution, since the number of points in view is random; in the basic case, a point has only a location, but more generally has a *mark*, a random variable in a general space.

A continuous-time Markov chain approach to simulating certain spatial point processes using birth-and-death processes was investigated by Preston (1977) and Ripley (1977).

– Geyer and Møller (1994) proposed a M-H sampler, as an alternative; their construction is a special case of reversible jump.

Stephens (2000): various trans-dimensional statistical problems can be viewed as abstract marked point processes.

He borrows the birth-and-death simulation idea to do finite mixture analysis, and also suggests that the approach appears to have much wider application, e.g. change point analysis and regression variable selection. The key feature of these three settings is the practicability of integrating out latent variables so that the likelihood is fully available.

Cappé, Robert and Rydén (2001) give a rather complete analysis of the relationship between reversible jump and continuous time birth-and-death samplers.

Product-space formulations

Several relatives of RJ work in a product space framework, in which the simulation keeps track of all θ_k, not only the 'current' one.

The state space is $\mathcal{K} \times \otimes_{k \in \mathcal{K}} \mathcal{R}^{n_k}$ instead of $\bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{R}^{n_k})$.

Advantage: circumvents the trans-dimensional character of the problem
Cost: requires 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).

Carlin and Chib (1995)

Let θ_{-k} denote 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

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

making the natural assumption that $p(Y|k, (\theta_l : l \in \mathcal{K})) = p(Y|k, \theta_k)$.

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 'pseudo-priors' is entirely a matter of convenience, but may influence sampler efficiency.

Carlin and Chib used conditionally independent pseudo-priors: $p(\theta_{-k}|k, \theta_k) = \prod_{l \neq k} p(\theta_l|k)$, 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: involves sampling from the pseudo-priors, so they design these pseudo-priors to ensure reasonable efficiency, by approximate matching to the posteriors: $p(\theta_l|k) \approx p(\theta_l|l, Y)$.

Variants on Carlin and Chib

Green and O'Hagan (1998) pointed out both that M-H moves could be made in this setting: also there is no need to update $\{\theta_l, l \neq k\}$ for irreducibility. In this form the pseudo-priors are only used in computing the update of k.

Dellaportas *et al.* (2002) proposed 'Metropolised Carlin and Chib' approach, in which joint model indicator / parameter updates were made: only necessary to resample the parameter vectors for the current and proposed models.

Composite model space framework

Godsill (2001) provides a general framework that embraces all of these methods, including reversible jump, facilitating comparisons between them. He takes a fixed pool of parameters $\{\theta_1, \theta_2, \ldots, \theta_N\}$, of which model k needs only $\theta_{\mathcal{I}(k)}$, parameter vectors that can overlap.

Then

$$p(k)p(\theta_{\mathcal{I}(k)}|k)p(\theta_{-\mathcal{I}(k)}|k, \theta_{\mathcal{I}(k)})p(Y|k, \theta_{\mathcal{I}(k)}),$$

The pseudo-prior is now $p(\theta_{-\mathcal{I}(k)}|k, \theta_{\mathcal{I}(k)})$.

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
- provides useful insight into some of the important factors governing the performance of reversible jump

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?

Alternatives to joint model-parameter sampling

The direct approach of an 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, 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 detail).

The posterior $p(\theta_k|Y,k)$ for the parameters θ_k is the target for an ordinary Bayesian MCMC calculation for model k.

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 factor is *Bayes factor* for model k_1 vs. k_0), to find $p(k|Y)$ 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.

Estimating marginal likelihoods

$$p(Y|k) = \left\{ \int [p(\theta_k|k,Y)/p(Y|k,\theta_k)] \mathrm{d}\theta_k \right\}^{-1}$$
$$= \int p(Y|k,\theta_k) p(\theta_k|k) \mathrm{d}\theta_k$$

so we have the estimates

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

and

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

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): 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 when 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.

Some issues in choosing a sampler

- Is k a model indicator really, or a parameter?
- Do we want results across k, within each k, or for one k of interest?
- Jumping between models as an aid to mixing (c.f. simulated tempering: mixing may be better in the 'other' model)
- Are samplers for individual models already written and tested?
- Are standard strategies like split / merge likely to work?
- Trade-off between remembering and forgetting θ_k when leaving model k

Methodological extensions

A simple automatic generic RJ sampler

For each model k, fix a n_k–vector μ_k and a $n_k \times n_k$–matrix B_k.

Suppose we are at (k, θ_k) and have proposed a move to model k', drawn from some transition matrix $(r_{k,k'})$.

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 trivial: 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 doesn't appear.

If the targets $p(\theta_k|k, y)$ were normal distributions, $N(\mu_k, B_k B_k^T)$, if the innovation variables u were $N(0, I)$, 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 M-H accept / reject decision. This is the motivation.

The idea might work adequately, if $p(\theta_k|k, y)$ are reasonably unimodal, with mean and variance approximately equal to μ_k and $B_k B_k^T$. Simple modifications:

- use t-distributions in place of the normals for u
- randomise over the orthogonal matrix R – or, to simplify implementation, take R to be a random permutation matrix
- use skewness transformations (David Hastie)
- use mixtures (Christophe Andrieu)

In practice, determine μ_k and B_k by short pilot runs within each k – only practical for a small finite set of models

Some experiments

These use a Fortran program, which calls a function written by the user to compute:

- $\log p(k, \theta_k, y)$
- 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.

(a) Variable selection in a small logistic regression problem

Dellaportas *et al.* (2002) illustrate their algorithm comparisons on a 2×2 factorial experiment with a binomially distributed response. All 5 interpretable models are entertained, with numbers of parameters (n_k) equal to 1, 2, 2, 3 and 4 respectively. We use the same 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 to be 2.90. The posterior model probabilities were computed to be (0.0051, 0.4929, 0.0113, 0.4388, 0.0519), 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. In this illustration, we condition on $1 \leq k \leq 6$. 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.

On this problem, the automatic sampler mixes much less well (presumably due to the extremely multi-modal parameter posteriors): the acceptance rate for model-jumping is 5.9%, while the integrated autocorrelation time rises to 118.

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. 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.

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 RJ setting by Green and Mira (2001), who call it *delayed rejection*.

If a proposal is rejected, instead of 'giving up', staying in the current state, and advancing time to the next transition, we instead attempt a second proposal, usually from a different distribution, and possibly dependent on the value of the rejected proposal.

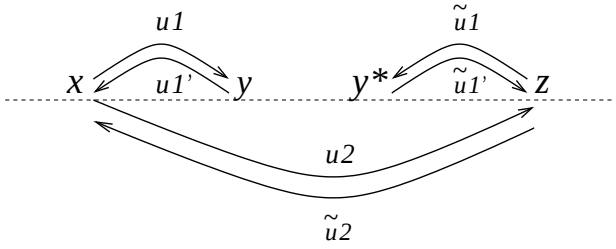
It is possible to set the acceptance probability for this second-stage proposal so that detailed balance is obtained, individually within each stage. The idea can be extended to further stages.

By the results of Peskun (1973) and Tierney (1998), this always reduces 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 in Green and Mira (2001) suggest that this can be the case.

The second-stage acceptance probability is calculated similarly as in deriving RJ above. We use two vectors of random numbers u_1 and u_2, drawn from g_1 and g_2, and two deterministic functions mapping these and the current state into the proposed new states, $y = h_1(x, u_1)$ and $z = h_2(x, u_1, u_2)$.

Both u_1 and u_2 appear in 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:

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

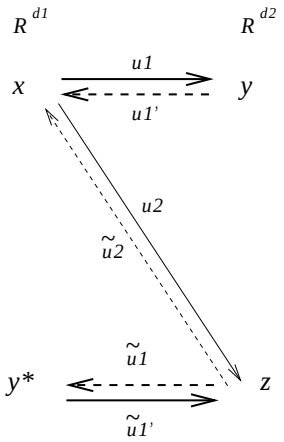
where u_1' is such that $x = h_1'(y, u_1')$.

Consider the case where the move to y is rejected. We need to find $\alpha_2(x, z)$ for detailed balance at the second-stage. As for one stage, 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 $y^\star = h_1'(z, \widetilde{u_1})$.

Equating integrands after making the change of variable, we find that a valid acceptance probability is

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

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



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'.

Efficient proposal choice for reversible jump MCMC

The most substantial recent methodological contribution to reversible jump MCMC generally is work by Brooks, Giudici and Roberts (RSS ordinary meeting, Banff, July 2002, *JRSS(B)*, 2002?) on the efficient construction of proposal distributions.

This is focussed mainly on the quantitative question of selecting the proposal density $g(u)$ 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 *et al.* propose several new methods, falling into two main classes.

1. using analysis of the acceptance rate as a function of u for small u (having chosen an appropriate scale of measurement for it), having assumed that uniformly high acceptance rate is desirable.
2. methods that work in a product-space formulation, 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.

References and preprints available from

http://www.stats.bris.ac.uk/∼peter

.../papers/hssschapter.ps

P.J.Green@bristol.ac.uk

Full written version: a chapter in the book *Highly Structured Stochastic Systems* (OUP, 2003).